UNIVERSIDADE FEDERAL DO AMAZONAS - UFAM
INSTITUTO DE CIÊNCIAS EXATAS - ICE
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA

AVALIAÇÃO DE CRITÉRIOS PARA A SELEÇÃO DO NÚMERO DE COMPONENTES EM MISTURAS FINITAS DE NORMAIS ASSIMÉTRICAS

JOSÉ MIR JUSTINO DA COSTA

MANAUS

2009

UNIVERSIDADE FEDERAL DO AMAZONAS - UFAM
INSTITUTO DE CIÊNCIAS EXATAS - ICE
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA

JOSÉ MIR JUSTINO DA COSTA

AVALIAÇÃO DE CRITÉRIOS PARA A SELEÇÃO DO NÚMERO DE COMPONENTES EM MISTURAS FINITAS DE NORMAIS ASSIMÉTRICAS

Dissertação apresentada ao Programa de Pós-Graduação em Matemática da Universidade Federal do Amazonas, como requisito parcial para obtenção do título de Mestre em Matemática, área de concentração em Estatística.

Orientador: Prof. Dr. José Raimundo Gomes Pereira

MANAUS
2009

Ao meu pai Antonino (in memorian)

JOSÉ MIR JUSTINO DA COSTA

AVALIAÇÃO DE ALGUNS CRITÉRIOS PARA A SELEÇÃO DO NÚMERO DE COMPONENTES EM MISTURAS FINITAS DE NORMAIS ASSIMÉTRICAS

Dissertação apresentada ao Programa de Pós-Graduação em Matemática da Universidade Federal do Amazonas, como requisito final para obtenção do título de Mestre em Matemática, área de concentração em Estatística.

Manaus, 17 de abril de 2009.

BANCA EXAMINADORA

.....

Prof. Dr. José Raimundo Gomes Pereira, Presidente

Universidade Federal do Amazonas-UFAM

.....

Prof. Dr. Celso Rômulo Barbosa Cabral, Membro

Universidade Federal do Amazonas-UFAM

.....

Prof. Dr. Victor Hugo Lachos, Membro

Universidade Estadual de Campinas-UNICAMP

AGRADECIMENTOS

Ao meu Deus, Pai de infinita bondade, através do qual encontro força e segurança para continuar a caminhada;

À meu pai Antonino (in memorian) e minha mãe D. Zefinha, pelo carinho dispensado durante toda a vida e pela crença nos estudos dos filhos como forma de superação dos problemas enfrentados;

À minha amada esposa Célia, companheira de todas horas que nunca me deixou abater nas dificuldades, que sempre acreditou nessa conquista e que dividiu cada angústia vivida nesta empreitada;

Aos meus filhos Jaqueline, Jean e Zé Lucas que souberam conviver com minha ausência e sempre serviram de motivação para continuar na luta;

Ao meu orientador Prof. José Raimundo que com sabedoria soube conduzir todos os passos desse trabalho;

Aos demais professores: Celso Rômulo (um verdadeiro co-orientador), José Cardoso, Amazoneida e Cícero pela gama de conhecimentos trasmitidos;

Ao prof. Victor Hugo Lachos e ao Rodrigo Basso por tanta contribuição e prestatividade;

Aos amigos Pe.John Docherty, Roberto Cristóvão, Valtemir, Domingos Anselmo e José Edson que acompanharam e torceram por cado novo passo dado em direção a conclusão desse trabalho e a todos os colegas do mestrado pela harmoniosa convivência, em especial à minha colega cearense, Leyne Marques, com quem partilhei horas de estudo e pelas suas importantes contribuições.

À FAPEAM pelo apoio financeiro.

RESUMO

AVALIAÇÃO DE CRITÉRIOS PARA A SELEÇÃO DO NÚMERO DE COMPONENTES EM MISTURAS FINITAS DE NORMAIS ASSIMÉTRICAS

Este trabalho tem por objetivo avaliar alguns critérios de informação para seleção de modelos no contexto de misturas finitas de normais assimétricas. Os critérios analisados foram o "Critério de Informação de Akaike-AIC", "Critério de Informação Bayesiano - BIC"e "Critério de Determinação Eficiente - EDC". A avaliação feita a respeito do desempenho apresentado por estes critérios se deu através de um estudo de simulação, em que utilizamos o algoritmo EM para encontrarmos as estimativas de máxima verossimilhança para os parâmetros do modelo com as quais empregamos os critérios. Foi também realizado uma aplicação da teoria desenvolvida para uma modelagem com dados reais utilizando dois conjuntos de dados já analisado anteriormente na literatura. Os resultados obtidos indicaram que, assintoticamente, os três critérios tendem a avaliar corretamente o número de componentes necessárias, mas para amostras pequenas o AIC apresenta desempenho inferior ao BIC e EDC, sendo que os dois últimos apresentam desempenho muito semelhante.

ABSTRACT

CRITERIA EVALUATION FOR THE SELECTION OF THE NUMBER OF COMPONENTS IN FINITE MIXTURES OF SKEW-NORMAL DISTRIBUTIONS

The present work aims to evaluate some information criteria for the selection of models in the context of finite mixtures of skew-normal distributions. The analyzed criteria are the Akaike's Information Criterion - AIC, the Bayesian Information Criterion - BIC and the Efficient Detection Criterion - EDC. The evaluation concerning the performance presented by these criteria was obtained through a simulation study, on which the EM algorithm is required to find the maximum likelihood estimates of for the parameters of the model where the criteria are applied. It was also performed an experiment for the application of the theory developed, modeling a real data set previously analyzed in the specific literature. The results obtained point that, in an asymptotic sense, the three criteria tend to correctly evaluate the number of necessary components, but for small samples the AIC presents inferior performance than BIC or EDC.

Sumário

Lista de Figuras

Lista de Tabelas

Introdução

Em muitas situações práticas, verificamos que a rotineira suposição de normalidade nem sempre é satisfeita, fato este que induz cada vez mais o uso de modelos mais flexíveis, como é o caso do modelo normal assimétrico. A distribuição normal assimétrica univariada foi introduzida por Azzalini (1985). Posteriormente, Azzalini & Dalla Vale (1996) estenderam a normal assimétrica ao caso multivariado. Azzalini & Capitanio (1999) enfatizaram aplicações estatísticas da versão multivariada.

No caso de interesse os dados são heterogêneos, apresentam multimodalidade e são provenientes de g populações distintas, mas não sabemos discriminá-las. Fazemos então essa modelagem através de misturas finitas de distribuições. A teoria inerente a esse contexto pode ser vista em McLachlan & Pell (2000). Não há dúvida que o modelo de mistura finita de densidades normais é o mais empregado nas aplicações que aparecem na literatura. Isso porque modelos de misturas finitas podem ser utilizados para representar densidades de qualquer complexidade, e de acordo com McLachlan & Pell (2000, seção. 6.1) qualquer distribuição pode ser aproximada, com uma precisão arbitrária por uma mistura finita de densidades normais. Como no passado a falta de métodos computacionais avançados tornava determinadas operações algébricas tediosas, então a simplicidade algébrica envolvida na distribuição normal se apresentava como uma grande vantagem. Embora esses modelos sejam atrativos, há ainda a necessidade de se checar as suposições distribucionais das componentes de mistura, pois além da heterogeneidade, em suas componentes os dados podem apresentar comporta-

mento assimétrico.

Lin *et al.* (2007) estendem a modelagem de misturas de normais usando misturas finitas de distribuições normais assimétricas univariadas, proposta por Azzalini (1985). Este modelo torna-se mais flexível ainda, em virtude da introdução de um novo parâmetro que regula a assimetria. Espera-se com isso que a modelagem de dados heterogêneos, multimodais e dotados de assimetria que utiliza misturas finitas de densidades normais assimétricas apresente resultados superiores àqueles que utilizam misturas finitas de normais. Apesar da introdução de um parâmetro a mais no modelo, o parâmetro que regula a assimetria, acredita-se que, nessa conjuntura, podemos modelar os dados com um número menor de componentes. Isso faz com que tenhamos uma menor quantidade de parâmetros a serem estimados para o modelo e também um ganho do ponto de vista computacional.

A motivação desse trabalho está no fato de muitos conjuntos de dados considerados na literatura apresentarem multimodalidade e comportamento não gaussiano, tais como assimetria. Portanto, o objetivo deste trabalho é avaliar o desempenho do Critério de Informação de Akaike-AIC, do Critério de Informação Bayesiano-BIC e do Critério de Determinação Eficiente-EDC para selecionar corretamente o número de componentes necessárias para realizar uma modelagem de dados usando uma mistura finita de densidades normais assimétricas. Os resultados do estudo de simulação realizado trazem em seu bojo informações que são importantes para profissionais das mais diversas áreas do conhecimento que lidem com análise de dados, em virtude de oferecer um estudo a cerca desses critérios que possibitam a realização de modelagem através de misturas finitas de densidades utilizando o menor número de componentes possível.

Esta dissertação está dividida em cinco capítulos. No Capítulo 1, apresentamos a distribuição normal assimétrica proposta por Azzalini (1985) e algumas definições e propriedades dessa distribuição para o caso univariado, os momentos

dessa distribuição além da sua representação hierárquica a qual usamos recorrentemente. Apresentamos também a estimação dos parâmetros do modelo via algoritmo EM.

A definição de misturas finitas de densidades, a estrutura de dados incompletos em modelos para problemas de misturas e como derivar o algoritmo EM para estimação de seus parâmetros estão discutidos no Capítulo 2.

No Capítulo 3 mostramos as definições e discussões dos critérios de informação utilizados nesse trabalho como métodos para seleção de modelos.

No Capítulo 4 apresentamos um estudo de simulação que é o cerne desse trabalho. Nele mostramos os procedimentos adotados , seus resultados e uma análise desses resultados.

No Capítulo 5 são apresentadas as conclusões acerca desse trabalho e algumas considerações.

Capítulo 1

O Modelo Normal Assimétrico

1.1 Definições e Propriedades

Neste capítulo será definido o modelo normal assimétrico sendo que algumas das vantagens dessa distribuição é que existem propriedades que são comuns com a distribuição normal além de ser bem mais tratável do ponto de vista matemático.

Apresentaremos também definições, proposições, lemas e propriedades necessárias ao desenvolvimento deste trabalho, onde algumas demonstrações podem ser vistas no apêndice A e mais detalhes encontra-se na literatura em Azzalini (1985).

Definição 1.1.1 *Dizemos que X tem distribuição normal assimétrica padrão com parâmetro* $\lambda \in \mathbb{R}$*, denotado por* $X \sim SN(\lambda)$*, se sua função densidade de probabilidade (fdp) é dada por*

$$g(x;\lambda) = 2\phi(x)\Phi(\lambda x), \quad -\infty < x < \infty \tag{1.1}$$

onde $\phi(.)$ *e* $\Phi(.)$ *são as funções de densidade de probabilidade e de distribuição da normal padrão, respectivamente.*

A notação SN que será usada de forma recorrente, para denotar normal assimétrica provém do inglês "skew normal".

A forma da distribuição é definida pelo parâmetro de assimetria λ. Valores positivos(negativos) de λ apontam uma assimetria positiva(negativa) no modelo, conforme podemos observar na figura (1.1) o comportamento da densidade para alguns valores de λ. Observe que a medida que $\lambda \to \infty$ ou $\lambda \to -\infty$, mais acentuada se torna a assimetria da distribuição.

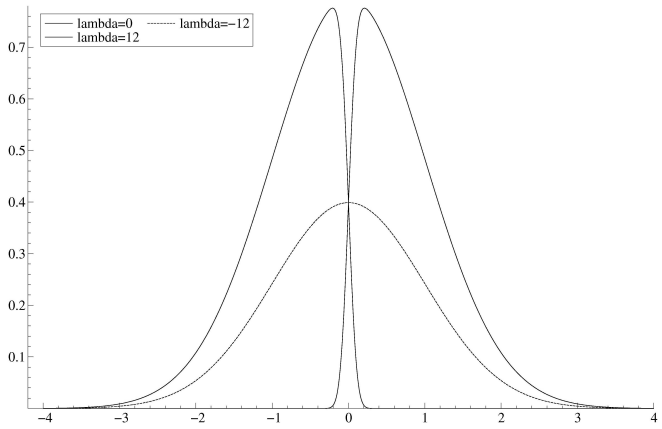

Figura 1.1: Função de densidade da distribuição normal assimétrica para $\lambda = 0, -12, 12$.

As propriedades que veremos a seguir são imediatas da definição (1.1.1):

Propriedade 1.1.1 *A função densidade de uma variável aleatória* $Y \sim SN(0)$ *é idêntica à de uma variável aleatória* $X \sim N(0,1)$

Propriedade 1.1.2 *Quando* $\lambda \to \infty$ *a densidade (1.1) tende para* $2\phi(x)I_{\{}x > 0\}$*, a qual corresponde à fdp de uma seminormal.*

Proposição 1.1.1 *Se* $Y \sim N(0,1)$ *e* $Z \sim SN(\lambda)$ *então* $\mid Z \mid$ *e* $\mid Y \mid$ *tem a mesma função densidade de probabilidade.*

Como consequência imediata da proposição (1.1.1) temos o seguinte resultado:

Propriedade 1.1.3 *Se $Y \sim SN(\lambda)$,então $Y^2 \sim \chi^2_1$.*

A propriedade a seguir nos fornece um importante método para simular observações de uma variável aleatória com fdp como dado em (1.1).

Propriedade 1.1.4 *Sejam Y e W variáveis aleatórias independentes e identicamente distribuídas com distribuição N(0,1), e definindo*

$$X = \begin{cases} Y, & \text{se} \quad \lambda Y > W, \\ -Y, & \text{se} \quad \lambda Y \leq W, \end{cases}$$

então $X \sim SN(\lambda)$.

Proposição 1.1.2 *Se (X,Y) é um vetor aleatório normal bivariado com distribuições marginais padronizadas e correlação δ, então a distribuição condicional de Y dado X >0 é $SN(\lambda(\delta))$.*

Escrevemos $SN(\lambda(\delta))$, para denotar que o parâmetro δ está relacionado à λ através da relação

$$\lambda(\delta) = \frac{\delta}{\sqrt{1-\delta^2}} \quad \text{ou} \quad \delta(\lambda) = \frac{\lambda}{\sqrt{1+\lambda^2}}. \tag{1.2}$$

O parâmetro δ varia no intervalo (-1,1) enquanto $\lambda \in \mathbb{R}$.
Doravante, por abuso de notação, vamos assumir $\delta(\lambda) = \delta$ e $\lambda(\delta) = \lambda$.

Proposição 1.1.3 *Se U_1 e U_2 são variáveis aleatórias independentes com distribuição normal padrão e $\delta \in (-1, 1)$. Então, $X \sim SN(\lambda)$ é tal que*

$$X = \delta|U_1| + \sqrt{1-\delta^2}U_2 \tag{1.3}$$

onde δ é definido em (1.2)

A Proposição 1.1.1 é muito útil na determinação dos momentos pares da distribuição normal assimétrica, dado que esses momentos são iguais aos de uma variável aleatória normalmente distribuída. Quanto aos momentos ímpares de uma variável aleatória $SN(\lambda)$, Henze (1986) apresenta a seguinte expressão, como consequência da Proposição 1.1.3:

$$E(X^{2k+1}) = \sqrt{\frac{2}{\pi}}\lambda(1+\lambda^2)^{-\frac{k+1}{2}}2^{-k}(2k+1)!\sum_{t=0}^{k}\frac{t!(2\lambda)^{2t}}{(2t+1)!(k-t)!}$$

O lema apresentado a seguir, dado em Azzalini (1985), é de suma importância para o cálculo da função geradora de momentos da normal assimétrica

Lema 1.1.1 *Se V tem distribuição $N(0,1)$, então,*

$$E[\Phi(hV+k] = \Phi\Big(\frac{k}{\sqrt{1+h^2}}\Big), \qquad \forall h,k \in \mathbb{R}$$

.

A função geradora de momentos de uma variável aleatória $Y \sim SN(\lambda)$ é dada por:

$$\begin{aligned}
M_Y(t) &= E\Big[e^{tY}\Big] = \int_{-\infty}^{\infty} e^{ty}2\phi(y)\Phi(\lambda y)dy \\
&= 2\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}+ty}\Phi(\lambda y)dy \\
&= 2\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(y^2-2ty+t^2\right)}e^{\frac{t^2}{2}}\Phi(\lambda y)dy \\
&= 2e^{\frac{t^2}{2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-t)^2}\Phi(\lambda y)dy; \qquad se \quad x = y-t \\
&= 2e^{\frac{t^2}{2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\Phi(\lambda x+\lambda t)dx \\
&= 2e^{\frac{t^2}{2}}E[\Phi(\lambda X+\lambda t)],
\end{aligned}$$

onde $X \sim N(0,1)$. Portanto, a partir de (1.1.1), temos que

$$M_Y(t) = 2e^{\frac{t^2}{2}}\Phi(\delta t), \tag{1.4}$$

onde

$$\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}.$$

Dessa forma de (1.4) podemos obter, após alguns cálculos, a esperança e a variância de Y e ainda os coeficientes de assimetria (γ_1) e curtose (γ_2) como dado abaixo:

$$E[Y] = b\delta, \tag{1.5}$$

$$Var[Y] = 1 - b^2\delta^2, \tag{1.6}$$

Onde,

$$b = \sqrt{\frac{2}{\pi}},$$

$$\gamma_1 = \frac{1}{2}(4-\pi)sinal(\lambda)\left[\frac{E^2(X)}{Var(X)}\right]^{\frac{3}{2}},$$

$$\gamma_2 = 2(\pi-3)\left[\frac{E^2(X)}{Var(X)}\right]^{2},$$

onde,

$$\begin{aligned}
\frac{E^2(X)}{Var(X)} &= \frac{\left(\left(\frac{2}{\pi}\right)^{\frac{1}{2}}\frac{\lambda}{\sqrt{1+\lambda^2}}\right)^2}{1-\frac{2}{\pi}\frac{\lambda^2}{1+\lambda^2}} \\
&= \frac{\frac{2}{\pi}\frac{\lambda^2}{1+\lambda^2}}{\frac{\pi(1+\lambda^2)-2\lambda^2}{\pi(1+\lambda^2)}} \\
&= \frac{2\lambda^2}{\pi(1+\lambda^2)-2\lambda^2} \\
&= \frac{\lambda^2}{\frac{\pi(1+\lambda^2)}{2}-\lambda^2} \\
&= \frac{\lambda^2}{\frac{\pi}{2}+\lambda^2\left(\frac{\pi}{2}-1\right)}.
\end{aligned} \tag{1.7}$$

Propriedade 1.1.5 *A função de distribuição acumulada de uma variável aleatória com fdp em (1.1) denotada por $F_Z(z;\lambda)$ é da forma*

$$F_Z(z;\lambda) = 2\Phi_2((z,0)^T;0,\Omega), \text{com } \Omega = \begin{bmatrix} 1 & -\delta \\ -\delta & 1 \end{bmatrix},\ \delta = \frac{\lambda}{\sqrt{1+\lambda^2}}. \tag{1.8}$$

onde,
$\Phi_2((z,0)^T;0,\Omega)$ *denota a função de distribuição acumulada de uma normal bivariada com vetor de médias zero e matriz de variância-covariância* $\mathbf{\Omega}$.

Demonstração:

$$
\begin{aligned}
F_Z(z;\lambda) &= 2\int_{-\infty}^{z}\phi(t)\Phi(\lambda t)dt = 2\int_{-\infty}^{z}\int_{-\infty}^{\lambda t}\phi(t)\phi(u)dudt \\
&\overset{(a)}{=} 2\sqrt{1+\lambda^2}\int_{-\infty}^{z}\int_{-\infty}^{0}\phi(t)\phi(v\sqrt{1+\lambda^2}+\lambda t)dvdt \\
&= 2\int_{-\infty}^{z}\int_{-\infty}^{0}\frac{\sqrt{1+\lambda^2}}{2\pi}\exp\left\{\frac{1}{2}\left(v^2(1+\lambda^2)+2\lambda\sqrt{1+\lambda^2}vt+\lambda^2t^2\right)\right\}dvdt \\
&= 2\int_{-\infty}^{z}\int_{-\infty}^{0}\frac{\sqrt{1+\lambda^2}}{2\pi}\exp\left\{\frac{1}{2}\left[\begin{pmatrix} t \\ v \end{pmatrix}^t\begin{pmatrix} 1+\lambda^2 & \lambda\sqrt{1+\lambda^2} \\ \lambda\sqrt{1+\lambda^2} & 1+\lambda^2 \end{pmatrix}\begin{pmatrix} t \\ v \end{pmatrix}\right]\right\}dvdt \\
&= 2\int_{-\infty}^{z}\int_{-\infty}^{0}\frac{\sqrt{1+\lambda^2}}{2\pi}\exp\left\{\frac{1}{2}\left[\begin{pmatrix} t \\ v \end{pmatrix}^t\begin{pmatrix} 1 & \frac{\lambda}{\sqrt{1+\lambda^2}} \\ \sqrt{1+\lambda^2} & 1 \end{pmatrix}^{-1}\begin{pmatrix} t \\ v \end{pmatrix}\right]\right\}dvdt \\
&= 2\int_{-\infty}^{z}\int_{-\infty}^{0}\frac{1}{2\pi\sqrt{1-\delta^2}}\exp\left\{\frac{1}{2}\left[\begin{pmatrix} t \\ v \end{pmatrix}^t\begin{pmatrix} 1 & -\delta \\ -\delta & 1 \end{pmatrix}^{-1}\begin{pmatrix} t \\ v \end{pmatrix}\right]\right\}dvdt,
\end{aligned}
\tag{1.9}
$$

onde (a) segue da transformação $v = \frac{u-\lambda t}{\sqrt{1+\lambda^2}}$

□

Propriedade 1.1.6 $1 - F_Z(-z;\lambda) = F_Z(z;-\lambda)$.

Propriedade 1.1.7 *Se* $Z \sim SN(\lambda)$, *então* $-Z \sim SN(-\lambda)$.

Propriedade 1.1.8 $F_Z(z;1) = [\Phi(z)]^2$.

A partir de $X \sim SN(\lambda)$, introduzimos os parâmetros ξ (de locação) e σ (de escala) via transformação $Y = \xi + \sigma X$, com $\xi \in \mathbb{R}$ e $\sigma > 0$, a qual conduz a seguinte definição:

Definição 1.1.2 *Dizemos que* Y *tem distribuição normal assimétrica com parâmetros de locação* ($\xi \in \mathbb{R}$), *de escala* ($\sigma^2 > 0$) *e assimetria* ($\lambda \in \mathbb{R}$), *denotado por* $Y \sim SN(\xi, \sigma^2, \lambda)$,*se sua fdp é dada por*

$$\psi(y;\xi,\sigma^2,\lambda) = \frac{2}{\sigma}\phi\Big(\frac{y-\xi}{\sigma}\Big)\Phi\Big(\lambda\frac{y-\xi}{\sigma}\Big), \tag{1.10}$$

Considerando a transformação $Y = \xi + \sigma X$ temos que se $Y \sim SN(\xi, \sigma^2, \lambda)$ então a sua função geradora de momentos é da forma:

$$\begin{aligned} M_Y(t) = E[e^{tY}] &= E[e^{t\,\xi + t\sigma X}] \\ &= e^{t\xi} E[e^{t\sigma X}] \\ &= e^{t\xi} M_X(t\sigma), \end{aligned} \tag{1.11}$$

onde $M_X(t\sigma)$ é a função geradora de momentos de X, dada em (1.4). Utilizando o Lema 1.1.1, encontramos

$$M_X(\sigma t) = 2e^{\frac{\sigma^2 t^2}{2}} \Phi\left\{\frac{\lambda\sigma\, t}{\sqrt{1+\lambda^2}}\right\} = 2e^{\frac{\sigma^2 t^2}{2}} \Phi(\delta\sigma t).$$

Substituindo $M_X(\sigma t)$ em (1.11), obtemos a função geradora de momentos de Y.

$$M_Y(t) = 2e^{\left(\xi\, t + \frac{\sigma^2 t^2}{2}\right)} \Phi(\delta\sigma t), \;\; t \in \mathbb{R}, \tag{1.12}$$

onde $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$

Assim, a esperança e a variância de Y é dada como segue

$$E[Y] = E[\xi + \sigma X] = \xi + \sigma E[X] = \xi + \sigma\delta\sqrt{\frac{2}{\pi}},$$

$$Var[Y] = Var[\xi + \sigma X] = \sigma^2 Var[X] = \sigma^2\left\{1 - \frac{2}{\pi}\delta^2\right\}.$$

1.2 Estimação dos Parâmetros via Algoritmo EM

Nesta seção empregaremos um algoritmo tipo EM, Dempster *et al.* (1977), para estimação de máxima verossimilhança dos parâmetros de uma distribuição normal assimétrica. Este algoritmo conhecido por ECM envolve dois passos: o Passo E(expectation) e o Passo CM(conditional maximization). Vale salientar

que o objetivo aqui é utilizar a teoria do algoritmo ECM para obtermos os estimadores de uma mistura finita de densidades normais assimétricas, que será visto no capítulo subsequente. Um resumo inerente a essa teoria pode ser vista no Apêndice C. Para empregar o algoritmo ECM representaremos o modelo normal assimétrico em uma estrutura de dados incompletos usando alguns resultados encontrados em Azzalini (1986, p. 201) e Henze (1986).

Lema 1.2.1 *(Representação Estocástica) Sejam T_0 e T_1 variáveis aleatórias independentes e com distribuição normal padrão. Então, se $X \sim SN(\lambda)$, segue-se que*

$$X = \delta|T_0| + (\sqrt{1+\delta^2})T_1. \tag{1.13}$$

Demonstração: (Basso, 2009)
Substituindo (1.13) na transformação $Y = \xi + \sigma X$, temos que

$$Y = \xi + \sigma(\delta|T_0| + \sqrt{1+\delta^2}T_1). \tag{1.14}$$

Para obtermos a densidade de Y empregaremos o método jacobiano. Sendo a inversa da transformação como dada abaixo

$$g^{-1}(y) = \frac{y - (\xi + \sigma\delta T_0)}{\sigma\sqrt{(1-\delta^2)}}$$

segue-se que a densidade de Y é dada por

$$f_Y(y) = f_{T_1}(g^{-1}(y))\left|\frac{\partial g^{-1}(y)}{\partial y}\right| = \frac{1}{\sigma\sqrt{(1-\delta^2)}\sqrt{2\pi}} e^{-\frac{(y-(\xi+\sigma\delta T_0))^2}{2\sigma^2(1-\delta^2)}} \tag{1.15}$$

Fazendo $T = \mid T_0 \mid$, temos então que a distribuição de Y dado $T = t$ é

$$Y|T = t \sim N(\xi + \sigma\delta T, (1-\delta^2)\sigma^2).$$

Logo, um modelo hieráquico para (1.14) pode ser escrito como

$$\begin{aligned} Y|T &\sim N(\xi + \sigma\delta T, (1-\delta^2)\sigma^2), \\ T &\sim N(0,1)I\{T>0\} \end{aligned} \tag{1.16}$$

Como na construção dos passos do algoritmo ECM nos deparamos com expressões não fechadas que dificultam a operacionalização do algoritmo usaremos a reparametrização proposta por Bayes & Branco (2007) cuja forma é a seguinte

$$\Gamma = (1-\delta^2)\sigma^2 \quad \text{e} \quad \Delta = \sigma\delta. \tag{1.17}$$

Precisamos, no entanto conhecer a distribuição de $f(t|y)$ para que possamos obter $E(T|Y=y)$ e $E(T^2|Y=y)$, ambas necessárias para a construção do Passo E do algoritmo.
Lembrando da definição de densidade condicional, temos que,

$$f(t|y) = \frac{f(y;t)}{\int_{-\infty}^{\infty} f(y;t)dt} = \frac{f(y|t)f(t)}{f(y)},$$

Com

$$\begin{aligned} f(t|y) &= \frac{1}{\pi\sqrt{1-\delta^2}\sigma}\exp(-\frac{t^2\delta^2\sigma^2}{2\sigma^2(1-\delta^2)})\exp(\frac{\delta\sigma t(y-\xi)}{\sigma^2(1-\delta^2)})\exp(-\frac{(y-\xi)^2}{2\sigma^2(1-\delta^2}})\exp(-\frac{t^2}{2}) \\ &\quad \frac{\sigma\sqrt{2\pi}}{2}\exp(\frac{(y-\xi)^2}{2\sigma^2})\Phi^{-1}(\lambda\eta) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1-\delta^2}}\exp(-\frac{t^2\sigma^2}{2\sigma^2(1-\delta^2)})\exp(\frac{2t\delta\sigma(y-\xi)}{2\sigma^2(1-\delta^2)})\exp(-\frac{(y-\xi)^2\delta^2}{2\sigma^2(1-\delta^2)})\Phi^{-1}(\lambda\eta) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1-\delta^2}}\exp(-\frac{(t\sigma-\delta(y-\xi))^2}{2\sigma^2(1-\delta^2)})\Phi^{-1}(\lambda\eta) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1-\delta^2}}\exp\Big(-\frac{1}{2}\Big(\frac{t-(\frac{\delta(y-\xi)}{\sigma})}{\sqrt{1-\delta^2}}\Big)^2\Big)\Phi^{-1}(\lambda\eta), \end{aligned} \tag{1.18}$$

onde

$$\mu_T = \frac{\delta(y-\xi)}{\sigma},\ \sigma_T^2 = (1-\delta^2) \text{ e } \eta = \frac{y-\xi}{\sigma}. \tag{1.19}$$

Portanto a distribuição condicional de T dado $Y = y,$ será

$$T|Y = y \quad \sim \quad TN(\mu_T, \sigma_T^2)\ \ I\{T > 0\} \tag{1.20}$$

Equivalentemente, se usarmos a reparametrização $\Gamma = (1-\delta^2)\sigma^2$ e $\Delta = \sigma\delta$, e

$$\mu_T = \frac{\Delta}{\Gamma + \Delta^2}(y-\xi) \quad \text{e} \quad \sigma_T^2 = \frac{\Gamma}{\Gamma + \Delta^2}, \tag{1.21}$$

temos que

$$\begin{aligned} f(t|y) &= \frac{1}{\sqrt{2\pi}(\sqrt{\Gamma})} e^{-\frac{1}{2\Gamma}(y-\xi-\Delta t)^2} \frac{2}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \left(\frac{2}{\sqrt{2\pi}\sigma}\right)^{-1} e^{\frac{1}{2}(\frac{y-\xi}{\sigma})^2} \Phi^{-1}(\lambda\eta) \\ &= \frac{1}{\pi\sqrt{\Gamma}} e^{-\frac{1}{2\Gamma}(y-\xi-\Delta t)^2} e^{-\frac{t^2}{2}} \Phi^{-1}(\lambda\eta), \end{aligned} \tag{1.22}$$

cuja distribuição condicional de T dado $Y = y$ é a mesma dada em (1.20).

Uma alternativa para facilitar o cálculo desta densidade condicional é usar Arellano-Valle *et al.* (2005). Agora, para que possamos determinar a função de log-verossimilhança dos dados aumentados, antes, vamos calcular densidade conjunta de Y e T, dada por:

$$\begin{aligned} f(y,t) &= f(y \mid t) f(t) \\ &= \frac{1}{\sqrt{2\pi}(\sqrt{\Gamma})} e^{-\frac{1}{2\Gamma}(y-\xi-\Delta t)^2} \frac{2}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \\ &= \frac{1}{\pi\sqrt{\Gamma}} e^{-\frac{1}{2\Gamma}(y-\xi-\Delta t)^2} e^{-\frac{t^2}{2}}. \end{aligned} \tag{1.23}$$

E considerando $Y_1, ..., Y_n$ e $T_1, ..., T_n$ amostras aleatórias para as variáveis definidas em (1.16) temos que a função de verossimilhança para os dados aumentados ou completos, denotada por $L(\Theta; y_i, t_i)$ é expressa como

$$\begin{aligned} L(\Theta; y_i, t_i) &= \prod_{i=1}^{n} \frac{1}{\pi\sqrt{\Gamma}} e^{-\frac{1}{2\Gamma}(y_i-\xi-\Delta t_i)^2} e^{-\frac{t_i^2}{2}} \\ &= \left(\frac{1}{\pi\sqrt{\Gamma}}\right)^n e^{-\frac{1}{2\Gamma}\sum_{i=1}^{n}(y_i-\xi-\Delta t_i)^2} e^{-\frac{1}{2}\sum_{i=1}^{n} t_i^2}. \end{aligned} \tag{1.24}$$

E tomando o $\log L(\Theta; y_i, t_i)$, obtemos a função de log-verossimilhança dos dados aumentados ou completos, denotada por $l_c(\Theta; y_i, t_i)$ como segue,

$$\begin{aligned} l_c(\Theta; y_i, t_i) &= n\log\left(\frac{1}{\pi\sqrt{\Gamma}}\right) - \frac{1}{2\Gamma}\sum_{i=1}^{n}(y_i-\xi-\Delta t_i)^2 - \frac{1}{2}\sum_{i=1}^{n} t_i^2 \\ &= n\left(-log(\pi\Gamma^{1/2})\right) - \frac{1}{2\Gamma}\sum_{i=1}^{n}(y_i-\xi-\Delta t_i)^2 - \frac{1}{2}\sum_{i+1}^{n} t_i^2 \\ &= -\frac{1}{2}\sum_{i=1}^{n} -n\log(\pi) - \frac{n}{2}\log(\Gamma) - \frac{1}{2\Gamma}\sum_{i=1}^{n}(y_i-\xi-\Delta t_i)^2 \\ &= c - \frac{n}{2}\log(\Gamma) - \frac{1}{2\Gamma}\sum_{i=1}^{n}(y_i-\xi-\Delta t_i)^2, \end{aligned} \tag{1.25}$$

onde $c = -\frac{1}{2}\sum_{i=1}^{n} t_i^2 - n\log(\pi)$ é uma constante que é independente do vetor de parâmetros $\boldsymbol{\Theta} = (\xi, \sigma^2, \lambda)$.

Como o passo E do algoritmo ECM, emprega as esperanças condicionais da função

$$Q(\boldsymbol{\Theta} \mid \hat{\boldsymbol{\Theta}}) = E_{\hat{\boldsymbol{\Theta}}}\Big[l_c(\boldsymbol{\Theta}) \mid y_i\Big], \tag{1.26}$$

torna-se necessário o Lema que enunciaremos a seguir extraído de Johnson *et al.* (1994):

Lema 1.2.2 *Sejam* $X \sim NT(\mu, \sigma^2)I\{a_1 < x < a_2\}$ *uma distribuição normal*

truncada com densidade dada por

$$f(y \mid \mu, \sigma^2) = \left\{ \Phi(\frac{a_2-\mu}{\sigma}) - \Phi(\frac{a_1-\mu}{\sigma}) \right\}^{-1} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}, a_1 < y < a_2. \tag{1.27}$$

Então

$$(i)E(X) = \mu - \sigma \frac{\phi(\alpha_2) - \phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)}, \tag{1.28}$$

$$(ii)E(X^2) = \mu^2 + \sigma^2 - \sigma^2 \frac{\alpha_2\phi(\alpha_2) - \phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)} - 2\mu\sigma \frac{\phi(\alpha_2) - \phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)}. \tag{1.29}$$

Prova: ver Apêndice B.

Do Lema 1.2.2 temos que

$$E(T_i|y_i) = \mu_{T_i} + \frac{\phi(\frac{\mu_{T_i}}{\sigma_T})}{\Phi(\frac{\mu_{T_i}}{\sigma_T})} \sigma_T \quad \text{e} \quad E(T_i^2|y_i) = \mu_{T_i}^2 + \sigma_T^2 + \frac{\phi(\frac{\mu_{T_i}}{\sigma_T})}{\Phi(\frac{\mu_{T_i}}{\sigma_T})} \mu_{T_i} \sigma_T \tag{1.30}$$

Sendo $\hat{s}_{1j}^{(k)} = E_{\hat{\Theta}^{(k)}}(T_i|y_i)$, $\hat{s}_{2j}^{(k)} = E_{\hat{\Theta}^{(k)}}(T_i^2|y_i)$ e usando as propriedades de esperança condicional conhecidas obtemos

$$\hat{s}_{1i}^{(k)} = \hat{\mu}_{T_i}^{(k)} + \frac{\phi\{\hat{\lambda}^{(k)}(\frac{y_i-\hat{\xi}^{(k)}}{\hat{\sigma}^{(k)}})\}}{\Phi\{\hat{\lambda}^{(k)}(\frac{y_i-\hat{\xi}^{(k)}}{\hat{\sigma}^{(k)}})\}} \hat{\sigma}_T^{(k)}, \tag{1.31}$$

$$\hat{s}_{2i}^{(k)} = \hat{\mu}_{T_i}^{2(k)} + \hat{\sigma}_T^{2(k)} + \frac{\phi\{\hat{\lambda}^{(k)}(\frac{y_i-\hat{\xi}^{(k)}}{\hat{\sigma}^{(k)}})\}}{\Phi\{\hat{\lambda}^{(k)}(\frac{y_i-\hat{\xi}^{(k)}}{\hat{\sigma}^{(k)}})\}} \hat{\mu}_{T_i}^{(k)} \hat{\sigma}_T^{(k)}, \tag{1.32}$$

onde, $\hat{\mu}_{T_i}^{(k)}$,$\hat{\sigma}_{T_i}^{(k)}$, são μ_{T_i} e σ_{T_i} dados em (1.21) com ξ, σ e λ substituidos por $\hat{\xi}^{(k)}$, $\hat{\sigma}^{(k)}$ e $\hat{\lambda}^{(k)}$, respectivamente.

Usando (1.20), (1.31) e (1.32) obtemos as esperanças condicionais dos dados completos dadas as observações, conforme mencionada em (1.26) , como sendo

$$\begin{aligned} Q(\boldsymbol{\Theta} \mid \hat{\boldsymbol{\Theta}}) &= E_{\hat{\boldsymbol{\Theta}}}\Big[l_c(\boldsymbol{\Theta}) \mid y_i\Big] \\ &= -\frac{n}{2}\log(\Gamma) - \frac{1}{2\Gamma}\sum_{i=1}^{n} y_i^2 - \frac{1}{2\Gamma}(n\Delta^2 \sum_{i=1}^{n} \hat{s}_{2i}) + \frac{2}{2\Gamma}(n\Delta \sum_{i=1}^{n}(y_i - \xi)\hat{s}_{1i}) + \\ &\quad + \frac{2}{2\Gamma}(n\xi \sum_{i=1}^{n} y_i) - \frac{n\xi^2}{2\Gamma}. \end{aligned} \tag{1.33}$$

A seguir apresentamos um resumo dos passos necessários para obtenção das estimativas de máxima verossimilhança para os parâmetros do modelo, sendo que mais detalhes podem ser visto no Apêndice C.

Passo E: Dado $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}$, calcule $\hat{s}_{1i}$ e $\hat{s}_{2i}$, para $i = 1, ..., n$.

Passo CM: Atualize $\hat{\Theta}^{(k)}$, maximizando $Q(\Theta|\hat{\Theta}) = E_{\hat{\Theta}}\{l_c(\Theta)|y_i\}$ e usando a reparametrização proposta por Arellano-Valle *et al.* (2005) após alguma álgebra nos leva as seguintes expressões fechadas:

$$\hat{\xi}^{(k+1)} = \sum_{i=1}^{n}(y_i - \Delta^{(k)}\hat{s}_{1i}^{(k)}), \tag{1.34}$$

$$\hat{\Gamma}^{(k+1)} = \frac{1}{n}\left\{\sum_{i=1}^{n}\left[(y_i - \hat{\xi}^{(k+1)})^2 - 2(y_i - \hat{\xi}^{(k+1)})\hat{\Delta}^{(k)}\hat{s}_{1i}^{(k)} - \hat{\Delta}^{2^{(k)}}\hat{s}_{2i}^{(k)}\right]\right\} \tag{1.35}$$

$$\hat{\Delta}^{k+1} = \frac{\sum_{i=1}^{n}(y_i - \hat{\xi}^{(k+1)})\hat{s}_{1i}^{(k)}}{\sum_{i=1}^{n}\hat{s}_{2i}^{(k)}} \tag{1.36}$$

As iterações do algoritmo são feitas através da repetição alternada dos passo E e CM sendo estas interrompidas quando um critério de convergência seja atingido. É comum utilizar a diferença $(l_c(\Theta^{(k+1)}) - l_c(\Theta^{(k)})$ o que significa que a convergência será atingida quando esta diferença for menor que uma constante c especificada, ou seja, o critério de parada adotado aqui foi

$$(l_c(\Theta^{(k+1)}) - l_c(\Theta^{(k)}) < c, \tag{1.37}$$

onde a constante c assumida neste trabalho é igual a 10^{-5}.

No próximo Capítulo apresentamos a definição de misturas finitas de densidades, com ênfase para a mistura finita de densidades normais assimétricas, o modelo de mistura em uma estrutura de dados incompletos e a derivação do algoritmo ECM para estimação dos parâmetros do modelo.

Capítulo 2

Misturas Finitas de Densidades

Neste capítulo definiremos o modelo de mistura finita de densidades o qual tem recebido muita atenção nos últimos anos por ser um modelo bastante flexível e em particular quando nos deparamos com situações onde há presença de heterogeneidade populacional.

Esses modelos tem aplicações em diversas áreas da estatística, como análise de agrupamento, análise discriminante e análise de sobrevivência, dentre outros. Existem vários trabalhos na literatura utilizando modelos de misturas para aproximar densidades complexas, desde aquelas com aspectos multimodais à outras totalmente assimétricas, sendo preferiveis em situações em que uma única familia paramétrica de distribuições não produz uma modelagem satisfatória. Será também definido o modelo de mistura finita de densidades normais assimétricas, sob o ponto de vista de dados incompletos e estimação dos parâmetros do modelo será feita pelo método da máxima verossimilhança via algoritmo ECM.

2.1 O Modelo de Mistura

Definição 2.1.1 *Um vetor aleatório* $\mathbf{Y} \in \mathbb{R}^p$ *com função densidade dada por*

$$f(\mathbf{y}) = \sum_{j=1}^{g} \omega_j \psi(\mathbf{y}), \tag{2.1}$$

onde, $\omega_j \geq 0$ *e* $\sum_{j=1}^{g} \omega_j = 1$*, é dito ter uma distribuição de mistura de densidades. A função* $f(\cdot)$ *é denominada mistura finita de densidades com g componentes, os parâmetros* $\omega_1, ..., \omega_g$ *são as proporções de misturas e as densidades* $\psi_1, ..., \psi_g$ *são as componentes da mistura.*

Se as componentes da mistura $\psi_j(\cdot)$ pertencem à famílias paramétricas de distribuições, então o modelo (2.1) pode ser reescrito da seguinte forma

$$f(\mathbf{y}; \boldsymbol{\Theta}) = \sum_{j=1}^{g} \omega_j \psi_j(\mathbf{y}; \theta_j), \tag{2.2}$$

onde $\boldsymbol{\Theta} = (\theta_1^T, ..., \theta_g^T)$ e θ_j são os parâmetros que definem cada uma das componentes ψ_j, que não precisam necessariamente estarem no mesmo espaço paramétrico. Mas para os propósitos desse trabalho assumiremos as componentes da mistura ψ_j como sendo pertecentes a mesma família paramétrica de distribuições o que nos leva a escrever a mistura finita de densidades por

$$f(\mathbf{y}; \boldsymbol{\Theta}) = \sum_{j=1}^{g} \omega_j \psi(\mathbf{y}; \theta_j), \mathbf{y} \in \mathbb{R}^p. \tag{2.3}$$

Com essas considerações temos que os parâmetros θ_j agora pertencem a um mesmo espaço parâmétrico.

2.2 A Mistura Finita de Densidades Normais Assimétricas

Considere uma amostra aleatória $Y_1, Y_2, ..., Y_n$ proveniente de mistura finita de densidades onde as componentes $\psi_j(\cdot)$ são densidades normais assimétricas. Temos então que um modelo de mistura de densidades normal assimétrica com g componentes é dado por

$$f(\mathbf{y} \mid \mathbf{\Theta}) = \sum_{j=1}^{g} \omega_j \psi(y_i \mid \xi_j, \sigma_j^2, \lambda_j), \tag{2.4}$$

onde $\omega = (\omega_1, ..., \omega_g)$ são as probabilidades da mistura, assumidas como não-negativa, cuja soma é um e $\mathbf{\Theta} = (\theta_1^T, ..., \theta_g^T)^T$ com $\theta_j = (\omega_j, \xi_j, \sigma_j^2, \lambda_j)^T$ sendo os parâmetros especificados para a componente i.

Para o contexto de modelagem por mistura finita de densidades introduziremos um conjunto de variáveis indicadoras latentes $Z_i = (Z_{i1}, ..., Z_{ig})^T, i = 1, ..., n$ cujos valores são um conjunto de variáveis binárias tais que

$$Z_{ik} = \begin{cases} 1 & \text{se } \ y_i \in k, \\ 0 & \text{caso contrário.} \end{cases}$$

e $\sum_{j=1}^{g} Z_{ij} = 1$.
Dadas as probabilidades da mistura $\omega = (\omega_1, ..., \omega_g)^T$ as componentes indicadoras $Z_1, ..., Z_n$ são independentes, com densidade multinomial dada por

$$f(z_i) = \omega_1^{z_{i1}} \omega_2^{z_{i2}} ...(1 - \omega_1 - ... - \omega_{g-1})^{z_{ig}}. \tag{2.5}$$

Escrevemos $Z_i \sim \mathcal{M}(1; \omega_1, ..., \omega_g)$ para denotar Z_i com densidade (2.5).

A partir da inclusão de Z_i faremos a modelagem dos dados ditos completos $y_c = (y_i, z_i)$. Assumindo novamente $(Y_1, ..., Y_n)$ amostras aleatórias da distribuição (2.4), temos que a densidade conjunta de $y_c = (y_i, z_i)$ é dada por:

$$f(y_c|\Theta) \ = \ f(y_i|z_i; \mathbf{\Theta}) f(z_i; \mathbf{\Theta}) \tag{2.6}$$

$$= \prod_{j=1}^{g}(f_j(y_i;\theta_j))^{(z_{ij})} f(z_i;\mathbf{\Theta}). \tag{2.7}$$

De acordo com a distribuição assumida por z_i em (2.5) temos que

$$f(z_i;\mathbf{\Theta}) = f(z_i;\omega) = \prod_{j=1}^{g}\omega_i^{(z_{ij})}, \tag{2.8}$$

e substituindo (2.8) em(2.6), temos

$$f(y_c|\Theta) = \prod_{j=1}^{g}\omega_i^{(z_{ij})}(f_j(y_i;\theta_j))^{(z_{ij})}. \tag{2.9}$$

Agora, considerando a independência dos dados incompletos e a expressão (2.9) temos que a função de verossimilhança é dada por:

$$\begin{aligned} L(\mathbf{\Theta}) &= \prod_{j=1}^{g} f(y_c|\Theta) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{g}\omega_i^{(z_{ij})}(f_j(y_i;\theta_j))^{(z_{ij})}, \end{aligned} \tag{2.10}$$

e, portanto, a função de log-verossimilhança dos dados completos é

$$l_c(\mathbf{\Theta}) = \sum_{i=1}^{n}\sum_{j=1}^{g} Z_{ij}\log\omega_i + \sum_{i=1}^{n}\sum_{j=1}^{g} Z_{ij}\log\left(f_j(y_i;\theta_j)\right). \tag{2.11}$$

Com essas considerações e assumindo a inclusão das variáveis indicadoras $Z_i^{'}s$ temos que um modelo hierárquico para uma mixtura finita de normais assimétricas pode ser escrito como

$$Y_i \mid t_i, Z_{ij} = 1 \sim N(\xi_j + \delta_j\sigma_j t_i, (1-\delta_j^2)\sigma_j^2), \tag{2.12}$$

$$T_i \mid Z_{ij} = 1 \sim NT(0,1)I\{T_i > 0\}, \tag{2.13}$$

$$Z_i \sim \mathcal{M}(1;\omega_1,...,\omega_g), \tag{2.14}$$

para $i = 1,...,n$, $j = 1,...,g$, $\delta_j = \frac{\lambda_j}{\sqrt{1+\lambda_j^2}}$ e NT denota a distribuição Normal Truncada.

De (1.20) temos que

$$T_i|y_i, Z_{ij} = 1 \sim NT(\mu_{T_{ij}}, \sigma_{T_j}^2)I\{T_i > 0\} \tag{2.15}$$

onde,

$$\mu_{T_{ij}} = \delta(y_i - \xi_j),\ \sigma_{Tj} = \sqrt{\Gamma_j} \tag{2.16}$$

Usando a reparametrização proposta por Arellano-Valle *et al.* (2005) e considerando $\mathbf{y} = (y_1^T, ..., y_n^T)$, $\mathbf{t} = (t_1, ..., t_n)^T$ e $Z = (Z_1^T, ..., Z_n^T)$ observações para as variáveis dadas em (2.12), (2.13) e (2.14), temos que a densidade conjunta baseada nas observações y_i é dada por

$$\begin{aligned} f(y_i, t_i, z_i) &= f(y_i|t_i, z_i) f(t_i|z_i) f(z_i) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{\Gamma_j}} e^{-\frac{(y_i - \xi - \Delta t_i)^2}{2\Gamma_j}} \frac{2}{\sqrt{2\pi}} e^{-\frac{t_i^2}{2}} \prod_{j=1}^{g} \omega_j^{z_{ij}} \\ &= \frac{1}{\pi\sqrt{\Gamma_j}} e^{-\frac{(y_i - \xi_j - \Delta t_i)^2}{2\Gamma_j}} e^{-\frac{t_i^2}{2}} \prod_{j=1}^{g} \omega_j^{z_{ij}} \end{aligned} \tag{2.17}$$

Dessa forma a funcão de verossimilhança para os dados completos denotada por $L_c(\Theta)$ é

$$\begin{aligned} L_c(\Theta) &= \prod_{i=1}^{n} f(y_c|\mathbf{\Theta}) \\ &= \prod_{i=1}^{n} f(y_i, t_i, z_i; \mathbf{\Theta}) f(t_i|z_i; \mathbf{\Theta}) f(z_i; \mathbf{\Theta}) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{g} \omega_j^{z_{ij}} \frac{1}{\pi\sqrt{\Gamma_j}} e^{-\frac{(y_i - \xi_j - \Delta t_i)^2}{2\Gamma_j}} e^{-\frac{t_i^2}{2}}. \end{aligned} \tag{2.18}$$

Agora, tomando o log de (2.18) a função de log-verossimilhança denotada por $l_c(\mathbf{\Theta})$ dos dados completos é

$$\begin{aligned} l_c(\mathbf{\Theta}) &= \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} \Big[\log \omega_j + \log f_i(y_i, t_i; \theta_i) \Big] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} \Big[\log \omega_j + \log \pi - \frac{1}{2} \log \Gamma_j - \frac{1}{2\Gamma_j} (y_i - \xi_j - \Delta_j t_i)^2 - \frac{t_i^2}{2} \Big] \\ &= c + \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} \Big[\log \omega_j - \frac{1}{2} \log \Gamma_j - \frac{1}{2\Gamma_j} (y_i - \xi_j - \Delta_j t_i)^2 \Big] \end{aligned} \tag{2.19}$$

onde $c = -\Big(\log \pi + \frac{t_i^2}{2} \Big)$ são constantes que não dependem do vetor de parâmetros

2.3 Estimação de Parâmetros Usando o Algoritmo tipo EM

Definiremos agora os passos necessários para a construção do algoritmo ECM, determinando a função Q e calculando as esperanças condicionais, como segue

$$\begin{aligned} Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(k)}) &= E_{\boldsymbol{\Theta}}\Big[l_c(\boldsymbol{\Theta}) \mid y_i\Big] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\log\omega_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\log\Gamma_j - \frac{1}{2\Gamma_j}\sum_{i=1}^{n}\sum_{j=1}^{k} E\Big[Z_{ij}Y_i^2|y_i\Big] - \\ &\quad -\frac{1}{2\Gamma_j}\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\xi_j^2 - \frac{1}{2\Gamma_j}\sum_{i=1}^{n}\sum_{j=1}^{k} E(Z_{ij}T_i^2\Delta_j^2 \mid y_i) + \\ &\quad +\frac{\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}Y_i\xi_i}{\Gamma_j} - \frac{1}{\Gamma}\sum_{i=1}^{n}\sum_{j=1}^{k} E(Z_{ij}T_i\xi_j\Delta_j \mid y_i) + \\ &\quad +\frac{1}{\Gamma_j}\sum_{i=1}^{n}\sum_{j=1}^{k} E(Z_{ij}T_i\Delta_j Y_i \mid y_i). \end{aligned} \tag{2.20}$$

Agora assumindo a independência de Z_{ij} obtemos as esperanças condicionais como definido abaixo:

$$\begin{aligned} \hat{z}_{ij}^{(k)} &= E[Z_{ij}|y_i;\hat{\Theta}^{(k)}], \\ \hat{s}_{1ij}^{(k)} &= E[Z_{ij}T_i|y_i;\hat{\Theta}^{(k)}], \\ \hat{s}_{2ij}^{(k)} &= E[Z_{ij}T_i^2|y_i;\hat{\Theta}^{(k)}], \end{aligned} \tag{2.21}$$

então podemos reescrever (2.20) da seguinte forma

$$\begin{aligned} Q(\boldsymbol{\Theta} \mid \hat{\boldsymbol{\Theta}}^{(k)}) &= \sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\log\omega_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\log\Gamma_j - \frac{1}{2\Gamma_j}\sum_{i=1}^{n}\sum_{j=1}^{k} y_i\hat{s}_{1ij}^{(k)} - \\ &\quad -\frac{1}{2\Gamma_j}\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\xi_j^2 - \frac{1}{2\Gamma}\sum_{i=1}^{n}\sum_{j=1}^{k} \Delta_j^2\hat{s}_{2ij}^{(k)} + \frac{1}{\Gamma}\sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}y_i\xi_j - \\ &\quad -\frac{1}{\Gamma}\sum_{i=1}^{n}\sum_{j=1}^{k} \xi_j\Delta_j\hat{s}_{1ij}^{(k)} + \frac{1}{\Gamma}\sum_{i=1}^{n}\sum_{j=1}^{k} y_i\Delta_j\hat{s}_{1ij}^{(k)}. \end{aligned} \tag{2.22}$$

Como no passo E do algoritmo ECM precisamos calcular as esperanças condicionais dadas em (2.21), o faremos como segue

$$\begin{aligned}\hat{z}_{ij} &= E[Z_{ij}|y_i;\hat{\Theta}^{(k)}],\\ &= E[Z_{ij}|\mathbf{y},\hat{\Theta}^{(k)}],\\ &= Pr[Z_{ij}=1|\mathbf{y},\hat{\Theta}^{(k)}],\end{aligned} \tag{2.23}$$

onde $\mathbf{y}=(y_1,y_2,...,y_n)$ são os dados observados.
Agora, assumindo a independência dos $z_i^{'s}$ e a distribuição dado por (2.14) podemos observar que

$$pr[Z_{ij}=1|\Theta^{(k)}]=Pr[Z_{ij}=1|z_{il}=0,\ \ \forall j\neq l|\theta^{(k)}]=\omega_i^{(k)}, \tag{2.24}$$

e usando (2.6), temos que

$$f(y_i|z_i;\Theta^{(k)})=\prod_{j=1}^{g}(f_j(y_i;\theta_j^{(k)}))^{(z_{ij})}, \tag{2.25}$$

de onde podemos observar que

$$f(y_i|Z_{ij}=1;\Theta^{(k)})=f_j(y_i;\theta_i^{(k)}), \tag{2.26}$$

E aplicando o Teorema de Bayes, temos

$$\hat{z}_{ij}=Pr[Z_{ij}=1|y_i,\Theta^{(k)}]=\frac{Pr[Z_{ij}=1|\Theta^{(k)}]f(y_i)|Z_{ij}=1;\Theta^{(k)}}{p(y_i;\Theta^{(k)})}, \tag{2.27}$$

Para concluir usaremos (2.4),(2.24) e (2.26) em (2.29) obtendo

$$\hat{z}_{ij}^{(k)}=\frac{\omega_j^{(k)}f_j(y_i;\theta_j^{(k)})}{\sum_{t=1}^{g}\omega_t^{(k)}f_t(y_i;\theta_t^{(k)})}, \tag{2.28}$$

logo,

$$\hat{z}_{ij}^{(k)}=\frac{\omega_j^{(k)}f(y_i|\omega_j^{(k)}\xi_j^{(k)},\sigma_j^{2^{(k)}},\lambda_j^{(k)})}{\psi(y_i|\Theta)^{(k)}}, \tag{2.29}$$

Verifica-se com isso que (2.29) é uma estimativa de propabilidade de y_i pertencer a uma população cuja distribuição seja dada por $f_j(.;\theta_j^{(k)})$.

Usando o Lema1.2.2 com $\Theta = \hat{\Theta}^{(k)}$ obtemos as demais esperanças condicionais necessárias ao passo E, como dado abaixo

$$\hat{s}_{1ij}^{(k)} = \hat{z}_{ij}^{(k)}\Big[\hat{\mu}_{T_{ij}}^{(k)} + \frac{\phi\{\hat{\lambda_j}^{(k)}(\frac{y_i-\hat{\xi_j}^{(k)}}{\hat{\sigma_j}^{(k)}})\}}{\Phi\{\hat{\lambda_j}^{(k)}(\frac{y_i-\hat{\xi_j}^{(k)}}{\hat{\sigma_j}^{(k)}})\}}\hat{\sigma}_{T_j}^{(k)}\Big], \tag{2.30}$$

$$\hat{s}_{2ij}^{(k)} = \hat{z}_{ij}^{(k)}\Big[\hat{\mu}_{T_{ij}}^{2(k)} + \hat{\sigma}_{T_j}^{2(k)} + \frac{\phi\{\hat{\lambda_j}^{(k)}(\frac{y_i-\hat{\xi_j}^{(k)}}{\hat{\sigma_j}^{(k)}})\}}{\Phi\{\hat{\lambda_j}^{(k)}(\frac{y_i-\hat{\xi_j}^{(k)}}{\hat{\sigma_j}^{(k)}})\}}\hat{\mu}_{T_{ij}}^{(k)}\hat{\sigma}_{T_j}^{(k)}\Big], \tag{2.31}$$

onde, $\hat{\mu}_{T_{ij}}^{(k)}$, $\hat{\sigma}_{T_j}^{(k)}$, são μ_{T_j} e σ_{T_j} dados em (2.16) com ξ, σ e λ substituidos por $\hat{\xi}^{(k)}$, $\hat{\sigma}^{(k)}$ e $\hat{\lambda}^{(k)}$, respectivamente.

No Passo CM do algoritmo fazemos as atualizações de $\hat{\Theta}^{(k)}$ através da maximização da função $Q(\Theta|\hat{\Theta}^{(k)})$. Como um modelo de mistura finita a atualização das estimativas $\hat{\omega_j}^{(k)}$ dos pesos $\omega_j's$ são determinadas através da maximização da função $Q(\Theta|\hat{\Theta}^{(k)})$ sujeito a restrição $\sum_{j=1}^{g}\omega_j = 1$, vamos fazê-lo reescrevendo a função Q e utilizando o método dos multiplicadores de Lagranje como dado abaixo

$$Q(\Theta|\Theta^{(k)}) = \sum_{i=1}^{n}\sum_{j=1}^{g}\hat{z}_{ij}^{(k)}\log\omega_j + \sum_{i=1}^{n}\sum_{j=1}^{g}\hat{z}_{ij}^{(k)}\log f_j(y_i;\theta_j) \tag{2.32}$$

$$= Q_1(\omega) + Q_2(\theta). \tag{2.33}$$

Escrevendo

$$M(\omega) = Q_1(\omega) + \zeta G(\omega), \tag{2.34}$$

onde $Q_1(\omega) = \sum_{i=1}^{n}\sum_{j=1}^{g}\hat{z}_{ij}^{(k)}\log\omega_j$ em (2.32) e $G(\omega) = 1 - \sum_{j=1}^{g}\omega_j$.
Agora, calculando as derivadas parciais em (2.34) com respeito a ω_j e ζ, respectivamente, temos:

$$\frac{\partial M(\omega)}{\partial\omega_j} = \frac{1}{\omega}\sum_{j=1}^{g}\hat{z}_{ij} - \zeta \quad \text{e} \quad \frac{\partial M(\omega)}{\partial\zeta} = 1 - \sum_{j=1}^{g}\omega_j, \tag{2.35}$$

igualando a zero cada uma das derivadas parciais em (2.35), obtemos

$$\left.\frac{\partial M(\omega)}{\partial \omega_j}\right|_{\omega_j=\hat{\omega}_j^{(k+1)}} = 0 \Rightarrow \hat{\omega}_j^{(k+1)} = \frac{1}{\zeta}\sum \hat{z}_{ij}^{(k)}, \tag{2.36}$$

$$\left.\frac{\partial M(\omega)}{\partial \omega_j}\right|_{\omega_j=\hat{\omega}_j^{(k+1)}} = 0 \Rightarrow \sum_{i=1}^{g} \hat{\omega}_j^{(k+1)} = 1. \tag{2.37}$$

Combinando (2.36) e (2.37), temos que,

$$1 = \sum_{i=1}^{g} \hat{\omega}_i^{(k+1)} = \frac{1}{\zeta}\sum_{j=1}^{n}\sum_{i=1}^{g} \hat{z}_{ij}^{(k)}. \tag{2.38}$$

Como $\sum\limits_{j=1}^{n}\sum\limits_{i=1}^{g} \hat{z}_{ij}^{(k)} = \sum\limits_{j=1}^{n} 1 = n$ em (2.38), temos que $\zeta = n$ e substituindo em (2.36) concluimos que

$$\hat{\omega}_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_{ij}^{(k)}. \tag{2.39}$$

A seguir apresentamos um resumo dos passos do algoritmo ECM para estimação dos parâmetros em um modelo de mistura normal assimétrica. Detalhes (ver Apêndice D)

Passo − E : Dado $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}^{(\mathbf{k})}$, calcule $\hat{z}_{ij}, \hat{s}_{1ij}, \hat{s}_{2ij}$, para $i = 1, ..., n$ e $j = 1, ..., g$.

Passo − CM : Atualize $\hat{\boldsymbol{\Theta}}^{(k)}$ maximizando $Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(k)}) = E_{\hat{\boldsymbol{\Theta}}}\Big[l_c(\boldsymbol{\Theta}) \mid y_i\Big]$ sobre $\boldsymbol{\Theta}$ de onde obtemos as seguintes expressões fechadas:

$$\hat{\omega}_j^{(k+1)} = \frac{1}{n}\sum_{j=1}^{n} \hat{z}_{ij}^{(k)}, \tag{2.40}$$

$$\hat{\xi}_j^{(k+1)} = \frac{\sum_{i=1}^{n}\left(\hat{z}_{ij}^{(k)} y_i - \hat{\Delta}_j \hat{s}_{1ij}^{(k)}\right)}{\sum_{i=1}^{n} \hat{s}_{1ij}^{(k)}}, \tag{2.41}$$

$$\hat{\Gamma}_j^{(k+1)} = \frac{1}{\sum\limits_{i=1}^{n} \hat{z}_{ij}^{(k)}}\left\{\sum_{i=1}^{n}\left(\hat{z}_{ij}^{(k)}(y_i - \hat{\xi}_j^{(k+1)})^2 - 2(y_i - \hat{\xi}_j^{(k+1)})\hat{\Delta}_j^{(k)}\hat{s}_{1ij^{(k)}} + \hat{\Delta}_j^{2^{(k)}}\hat{s}_{2ij}^{(k)}\right)\right\}, \tag{2.42}$$

$$\hat{\Delta}_j^{(k+1)} = \frac{\sum_{i=1}^{n}\left(y_i - \xi_j^{(k+1)}\right)\hat{s}_{1ij}^{(k)}}{\sum_{i=1}^{n} \hat{s}_{2ij}^{(k)}}, \tag{2.43}$$

Capítulo 3

Estimação do Número de Componentes

Um dos problemas enfretados com aplicações de mistura finita de densidades é estimar o número de componentes e consequentemente, o número de parâmetros adequado para os dados. Portanto, a idéia subjacente desta abordagem é estabelecer uma função-critério que sinalize o verdadeiro número de componentes do modelo. Normalmente são da forma

$$\hat{g} = \arg min_g \{C(\hat{\Theta}_{(g)}), g = g_{min}, ..., g_{max}\}, \tag{3.1}$$

onde $C(\hat{\Theta}_{(g)})$ é o valor da função-critério para o modelo estimado com dimensão g. Serão avaliados três critérios de informação, *Critério de Informação de Akaike-AIC, Critério de Informação Bayesiano-BIC e Critério de Determinação Eficiente-EDC*, com o objetivo de obtermos as estimativas de g. O AIC e o BIC são critérios amplamente conhecidos na literatura estatística enquanto o EDC surge como uma nova alternativa para estudo de selecão de modelos. Foi desenvolvido originalmente por Bai *et al.* (1989) para detectar a quantidade de sinais na presença de ruído. Em Zhao *et al.* (2001) este critério é novamente utilizado para selecionar a ordem de uma cadeia de Markov.

É importante ressaltar que esses critérios não devem ser utilizados como regras de decisões, mas, como ferramenta que dá evidências de quantas componentes

deve ser usada em detrimento de outras. Uma discussão mais aprofundada sobre este assunto pode ser vista em McLachlan & Pell (2000, Seção 6.8).

3.1 O Critério de Informação de Akaike

A idéia do Critério de Informação de Akaike - AIC(*Akaike's Information Criterion*) é estimar a informação de Kullback & Leibler (1951) do verdadeiro modelo com respeito ao modelo estimado, cuja forma é:

$$\begin{aligned} I_{KL}(\boldsymbol{\Theta}^*|\hat{\boldsymbol{\Theta}}) &= \int f(\mathbf{x}|\boldsymbol{\Theta}^*) \log (\frac{f(\mathbf{x}|\boldsymbol{\Theta}^*)}{f(\mathbf{x}|\hat{\boldsymbol{\Theta}})}) d\mathbf{x} \\ &= \int f(\mathbf{x}|\boldsymbol{\Theta}^*) \log f(\mathbf{x}|\boldsymbol{\Theta}^*) d\mathbf{x} - \int f(\mathbf{x}|\boldsymbol{\Theta}^*) \log f(\mathbf{x}|\hat{\boldsymbol{\Theta}}) d\mathbf{x}, \quad (3.2) \end{aligned}$$

onde $f(\mathbf{x}|\boldsymbol{\Theta}^*)$ denota a verdadeira densidade e $f(\mathbf{x}|\hat{\boldsymbol{\Theta}})$ o modelo estimado.

Como a informação de Kullback-Leibler mede a divergência entre o modelo verdadeiro e o modelo estimado, o objetivo portanto passa ser minimizar essa divergência. Como o primeiro termo no lado direito de (3.2) não depende do modelo estimado, somente o segundo termo é relevante à minimização. Assumindo F como a verdadeira distribuição e $\mathbf{y} = (y_1^T, ..., y_n^T)$ os dados observados, obtemos o que chamamos de log-verossimilhança esperada como dada abaixo

$$\begin{aligned} \eta(\mathbf{y}; F) &= \int f(\mathbf{x}|\boldsymbol{\Theta}^*) \log f(\mathbf{x}|\hat{\boldsymbol{\Theta}}) d\mathbf{x} \\ &= \int \log f(\mathbf{x}|\hat{\boldsymbol{\Theta}}) dF(\mathbf{x}), \quad (3.3) \end{aligned}$$

Na derivação do AIC, $\eta(\mathbf{y}; F)$ é estimada por

$$\eta(\mathbf{y}; \hat{F}_n) = \frac{1}{n} \sum_{i=1}^{n} \log f(\mathbf{y}_i|\hat{\boldsymbol{\Theta}}), \quad (3.4)$$

obtido substituindo F em (3.3) pela sua distribuição empírica, a qual atribui massa $1/n$ em cada observação $\mathbf{y}_i$, $(i = 1, ..., n)$. Mas, em geral isso produz uma superestimativa da log-verossimilhança média, uma vez que que a função de distribuição empírica é geralmente mais próxima da distribuição estimada do que

da verdadeira distribuição desconhecida. Então o viés do estimador $\hat{\eta}(\mathbf{y}|\hat{\Theta})$ é o funcional

$$\begin{aligned} b(F) &= E_F[\eta(\mathbf{Y};\hat{F}_n) - \eta(\mathbf{Y};F)] \\ &= E_F\Big[\frac{1}{n}\sum_{i=1}^{n}\log f(\mathbf{y}_i|\hat{\mathbf{\Theta}}) - \int \log f(\mathbf{x}|\hat{\mathbf{\Theta}})dF(\mathbf{x})\Big]. \end{aligned} \tag{3.5}$$

Um critério de informação para seleção de modelos pode ser construído baseado na log-verossimilhança e levando em consideração a correção desse viés. Com essas considerações podemos assumí-lo da seguinte forma,

$$\log L(\hat{\mathbf{\Theta}}) - b(F), \tag{3.6}$$

no qual devemos usar uma estimativa apropriada para o termo $b(F)$. O que se pretende com isso é escolher o modelo mais adequado para que seja maximizada a relação (3.6). Na literatura, o critério de informação normalmente é dado pelo dobro do valor negativo dessa diferença, então assume a forma

$$-2\log L(\hat{\Theta}) + C, \tag{3.7}$$

onde o primeiro termo da relação anterior mede a falta de ajuste do modelo e o segundo termo C é uma penalização que mede a complexidade do modelo. Assim sendo, (3.7) objetiva escolher o modelo que minimize esse critério.

Em Akaike (1974) foi mostrado que $b(F)$ é assintoticamente igual ao número de parâmetros livres a ser estimado no modelo de dimensão g, o qual denotou por d. Portanto, o critério de informação de Akaike seleciona o modelo que minimiza

$$AIC(g) = -2L(\hat{\Theta}_{(g)}) + 2d_{(g)}. \tag{3.8}$$

Acontece que do ponto de vista teórico o AIC assume o modelo verdadeiro e o conjecturado como sendo pertencentes a uma mesma família paramétrica de distribuições e mais, assume as condicões de regularidade da teoria assintótica

que não se verificam no contexto de misturas finitas.

Embora continue sendo um dos critérios ainda bastante utilizado na literatura para seleção de modelos, em trabalhos realizados por vários autores, foi verificado que o AIC não é consistente em ordem e tende a superestimar a dimensão do modelo, ver por exemplo, Celeux & Soromenho (1996). No contexto de misturas isto significa que o AIC tende a selecionar o modelo com um número maior de componentes que o verdadeiro.

Mais detalhes do desenvolvimento para a determinação do AIC no contexto de misturas finitas de densidades pode ser visto em McLachlan & Pell (2000, Seção 6.8).

3.2 Critério de Informação Bayesiano - BIC

O Critério de Informação Bayesiano - BIC (*Bayesian Information Criterion*) está baseado na teoria Bayesiana de seleção de modelos. Nessa teoria, são considerados vários possíveis modelos, com suas probabilidades a priori, e o objetivo é selecionar o modelo com a maior probabilidade a posteriori dadas as observações $\mathbf{y} = y_1, ..., y_n$. Sendo $M_1, M_2, ..., M_g$ os modelos considerados e $p(M_g)$, $g = 1, ..., G$, as respectivas probabilidades a priori, pelo Teorema de Bayes, a posteriori de M_g dado $\mathbf{y}$ é

$$p(M_g \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid M_g)p(M_g)}{\sum_{t=1}^{g} p(\mathbf{y} \mid M_t)p(M_t)} \tag{3.9}$$

De (3.9) vemos que, para a posteriori de M_g, é necessário determinar $p(\mathbf{y} \mid M_g)$. Quando existem parâmetros desconhecidos nos modelos, essa distribuição é obtida por integração sobre o espaço dos parâmetros, ou seja,

$$p(\mathbf{y} \mid M_g) = \int p(\mathbf{y} \mid \Theta_{(g)}, M_g)p(\Theta_{(g)} \mid M_g)d\Theta_{(g)} \tag{3.10}$$

onde $p(\Theta_{(g)} \mid M_g)$ é a distribuição a priori para $\Theta_{(g)}$, ver Kass & Raftery (1995). Vale ressaltar que $p(\mathbf{y} \mid \Theta_{(g)}, M_g)$ é a função de verossimilhança para o modelo

M_g, com vetor de parâmetros $\Theta_{(g)}$.

A quantidade $p(\mathbf{y} \mid M_g)$ recebe a denominação de verossimilhança integrada. Se as probabilidades a priori $p(M_{(g)})$forem todas iguais, o procedimento seleciona o modelo com a maior verossimilhança integrada.

A maior dificuldade com essa abordagem Bayesiana é avaliar a integral que define a verossimilhança integrada. A principal abordagem para aproximar essa integral é através da minimização do Critério de Informação Bayesiano, que é dado por,

$$BIC(g) = -2 \log L(\hat{\Theta}) + d_{(g)} \log n, \tag{3.11}$$

onde $d_{(g)}$ é o número de parâmetros a ser estimado no modelo de dimensão g. Este critério foi desenvolvido por Schwarz (1978) e por isso também é conhecido na literatura por *Critério de Schwarz*.

O BIC é considerado consistente em ordem, o que implica que assintoticamente tende a selecionar o modelo de dimensão correta ver Celeux & Soromenho (1996). Esse critério foi desenvolvido sob condições de regularidade que também não se verificam para modelos de mistura finita. No entanto, em trabalhos realizados, seu emprego tem demonstrado resultados bastante satisfatórios. Em particular, esses bons resultados têm se verificado empregando-se esse critério para selecionar o número de componentes para uma mistura finita em estimação de densidades, ver Biernacki *et al.* (2000).

3.3 Critério de Determinação Eficiente - EDC

O Critério de Determinação Eficiente - EDC (*Efficient Determination criterion*) idealizado por Bai *et al.* (1989) foi utilizado com o objetivo de detectar o número de sinais transmitidos na presença de ruidos, um problema comum em processamento de sinais. Em Zhao *et al.* (2001) novamente este critério foi usado, sendo que desta vez para avaliar a ordem de uma Cadeia de Markov com espaço

de estado finito. Este critério generaliza os critérios AIC e BIC e produz uma classe de estimadores consistentes para a ordem da cadeia acima mencionada. Neste trabalho usaremos o EDC com a finalidade de investigarmos se os bons resultados apresentados nos trabalhos citados podem ser estendidos para avaliar o número de componentes envolvidas em um modelo de mistura finita de normais assimétricas. Neste contexto o critério proposto por Zhao *et al.* (2001) é dado por

$$EDC(g) = -2\log L(\hat{\Theta}) + d_{(g)}c_n, \tag{3.12}$$

onde c_n no termo de penalização, pode ser tomado como uma sequência de números positivos dependendo de n, onde em nosso estudo, n é o tamanho da amostra. Aqui $d_{(g)}$ será assumido como o número de parâmetros a ser estimado no modelo de dimensão g. A vantagem deste critério é que podemos flexibilizar a escolha de c_n, precisando apenas que as seguintes condições sejam satisfeitas:

1. $\frac{c_n}{n} \to 0$ quase certamente quando n tende ao infinito.
2. $\frac{c_n}{\log\log n} \to \infty$ quase certamente quando n tende ao infinito.

O problema que enfrentamos, é o de escolher o c_n que produza resultados mais eficientes, haja vista que podemos escolhê-lo de uma classe de funções que pode ser infinita. No entanto, em Zhao *et al.* (2001) encontramos duas sugestões para escolha do c_n, sendo $\{c_1, c_2\}$ como dado abaixo, constantes escolhidas apropriadamente.
$c_n^{(1)} = c_1 n/\log n$ e $c_n^{(2)} = c_2\sqrt{n}$,

No próximo capítulo apresentaremos os resultados do estudo comparativo realizado, utilizandos os critérios AIC, BIC e EDC.

Capítulo 4

Estudo de Simulação e Aplicações

Neste capítulo apresentamos um estudo de simulação que visa analisar o desempenho dos critérios de informação AIC, BIC e EDC para selecionar corretamente o número de componentes necessárias para modelar um conjunto de dados através de uma mistura finita de densidades normais assimétricas.

Os gráficos apresentados ao longo deste trabalho bem como as simulações e análise de dados foram todos confeccionados no ambiente de programação R Development Core Team (2004) em sua versão 2.7.0. R é distribuída gratuitamente e está disponível em http://www.r-project.org. Uma grande vantagem dessa linguagem é o fato de ser utilizada amplamente no meio acadêmico, o que contribui para a imensa variedade de pacotes desenvolvidos nas diversas áreas da estatística.

4.1 Descrição do Experimento

No intuito de verificarmos se os critérios realmente sinalizavam para o número correto de componentes envolvidos num modelo , inicialmente adotamos os seguintes procedimentos:

1. geramos amostras de tamanho 200, 300, 500 e 1000 sendo todas proveniente de uma mistura de $g = 3$ componentes normais assimétricas com os

seguintes parâmetros, $(\omega_1 = \omega_2 = \omega_3 = 1/3, \xi_1 = 5, \xi_2 = 20, \xi_3 = 28, \sigma_1^2 = 9, \sigma_2^2 = 16, \sigma_3^2 = 16, \lambda_1 = 6, \lambda_2 = -4, \lambda_3 = 4)$. As Figuras 4.1 e 4.2 ilustram a distribuição desses dados com os parâmetros assumidos para dois diferentes tamanhos de amostras.

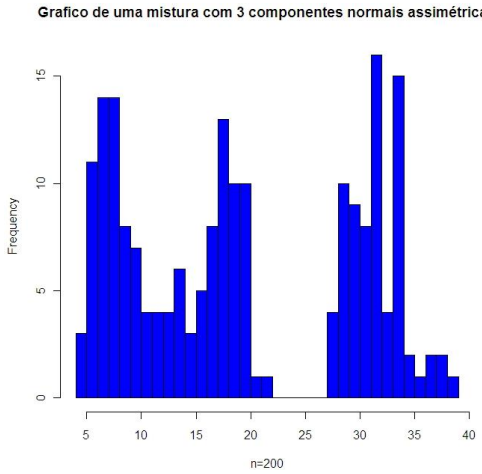

Figura 4.1: Mistura de 3 componentes SN com $n = 200$.

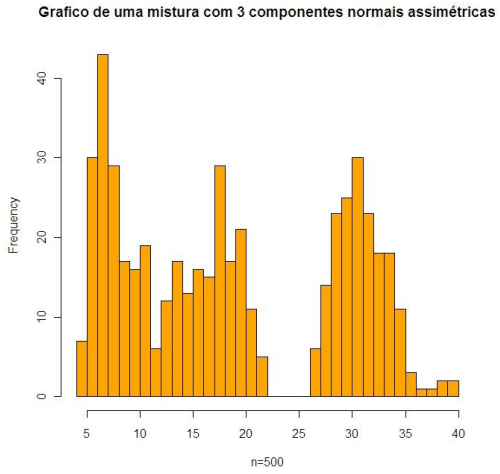

Figura 4.2: Mistura de 3 componentes SN com $n = 500$.

2. dividimos cada amostra gerada em g grupos, $g = \{2, 3, 4, 5\}$, usando o método k-means, ver Wichern & Johnson (2007). Para inicializarmos o algoritmo EM, utilizamos o método dos momentos, sendo os parâmetros de locação, escala e assimetria calculados através das expressões abaixo extraidas de Lin *et al.* (2007)

$$\xi_g = \mu_g - \sqrt{\frac{2}{\pi}}\delta\sigma, \tag{4.1}$$

$$\sigma_g^2 = \frac{\omega_g}{1 - \frac{2}{\pi}\delta^2}, \tag{4.2}$$

$$\lambda_g = \pm\sqrt{\frac{\pi\gamma_g^{2/3}}{2^{1/3}(4-\pi)^{2/3} - (\pi-2)\gamma_g^{2/3}}} \tag{4.3}$$

onde, γ_g é o coeficiente de assimetria, cujos valores estão no intervalo (-0.9953, 0.9953) e $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$, ver Lin *et al.* (2007).

Para cada grupo, dividido por este método, teremos como estimativas iniciais da média, da variância e do coeficiente de assimetria, a média amostral, a variância amostral e o coeficiente de assimetria amostral, respectivamente. Com relação à estimativa inicial para o peso será usado

$$\hat{\omega}_g^{(0)} = \frac{n_g}{n} \quad \text{e} \quad \hat{\omega}_{g+1}^{(0)} = 1 - \hat{\omega}_g^{(0)}$$

onde n_g representa a g-ésima parte da amostra.

3. Após submetermos ao algoritmo EM cada amostra de acordo com os g grupos definidos, obtivemos a log-verossimilhança estimada para uma mistura com $g = 2$, $g = 3$, $g = 4$ e $g = 5$ componentes. Com base nelas calculamos o AIC, o BIC e o EDC, registrando seus resultados.

   A condição de parada imposta ao algoritmo EM neste trabalho, foi baseado na diferença entre as estimativas sucessivas da função de log-verossimilhança, isto é,

   $$|l_c(\theta^{(k+1)}) - l_c(\theta^{(k)})| < 10^{-5}$$

   onde,
   $l_c(\theta^{\cdot})$ denota a log-verossimilhança estimada.

4. Geramos novas amostras com os mesmos tamanhos já relatados, com os mesmos parâmetros e seguindo os mesmos procedimentos anteriores. Registrou-se novamente o resultado produzido pelos critérios.

5. Repetimos este experimento 500 vezes, e determinamos o percentual de acerto dos critérios para cada tamanho de amostra analisado nestas repetições.

Como para o cálculo dos critérios somente o termo de "penalizacão"do EDC pode sofrer modificacões na sua forma funcional em virtude da flexibilização do c_n, então decidimos avaliar diferentes formas para a função c_n, considerando-os como dado abaixo:

1. $c_n = 0.2\sqrt{n}$

2. $c_n = 0.2 \log n$

3. $c_n = 0.2n/\log n$

4. $c_n = 0.3\sqrt{n}$

Vale ressaltar que o logaritmo aqui assumido é o logaritmo neperiano.

4.2 Análise dos Resultados

Os resultados deste estudo de simulação podem vistos através das tabelas e figuras constantes nesta seção. A Tabela 4.1 e o Gráfico 4.3 mostram a quantidade de vezes, em valores percentuais, que cada critério indicou corretamente para $g = 3$ componentes, em 500 repetições do experimento, assumindo assumindo as diferentes formas da função c_n.

Percebe-se na Tabela 4.1 que o desempenho do AIC é inferior ao dos demais critérios para todos os tamanhos de amostra estudado. Talvez possamos atribuir esta performance inferior ao fato do AIC não levar em conta o tamanho da amostra no "termo de penalização", apenas o número de parâmetros a ser estimado no modelo. Isto faz com que a log-verossimilhança estimada seja mais fortemente penalizada pelo BIC e EDC que consideram o tamanho da amostra em seus termos de "penalização". Observe que para amostras de tamanho 200 o percentual de acerto do AIC para $g = 3$ componentes está em torno de oitenta e sete por cento, portanto, inferior aos resultados produzidos pelo BIC e EDC (com $c_n = 0.2\sqrt{n}$).

Tamanho da amostra	**AIC**	**BIC**	**EDC**			
			$c_n = 0.2\sqrt{n}$	$c_n = 0.2 \log n$	$c_n = 0.2n/\log n$	$c_n = 0.3\sqrt{n}$
200	87.7	98	98	66.1	98	98
300	91.7	99.3	99.3	75.6	99.6	99.3
500	92.8	100	100	83.3	100	100
1000	94.2	100	100	85	100	100

Tabela 4.1: Desempenho dos Critérios de Informação em % de acerto.

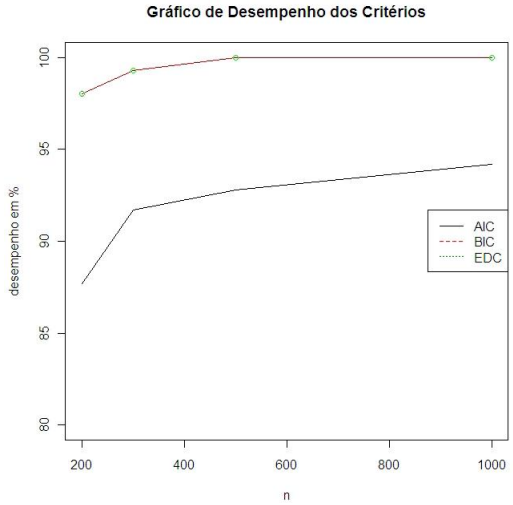

Figura 4.3: Gráfico de Desempenho dos Critérios AIC, BIC e EDC com $c_n = 0.2\sqrt{n}$.

Quanto ao BIC, os resultados mostrados na Tabela 4.1, confirmam os bons resultados já apresentados em outros trabalhos na literatura estatística. Mesmo para amostras de tamanho 200 e 300, que são as menores analisadas neste estudo o seu desempenho para sinalizar corretamente o número de componentes atingiu o patamar de 98% e 99.3% de acertos, respectivamente. O fato do BIC considerar o tamanho da amostra, faz com que o termo de "penalização"seja sempre maior que o do AIC. Como estamos considerando o logaritmo neperiano, mesmo para amostras de tamanho n=10, por exemplo, o segundo termo em (3.11) penaliza mais fortemente a log-verossimilhança do que no AIC. É possível que este fato

faça com que a divergência entre o modelo verdadeiro e o modelo estimado seja minimizada de forma eficiente pelo BIC com relação ao AIC.

Já em relação ao EDC, de acordo com o que se pode observar na Tabela 4.1 e nas figuras 4.4, 4.5 e 4.6 uma análise comparativa deve sempre atentar para a questão do c_n assumido. Ou seja, se a escolha do c_n não for adequada para os dados em análise podemos incorrer no erro de obtermos um resultado inferior ao que o critério poderia produzir. As demais funções c_n utilizadas neste trabalho foram as seguintes:

- $c_n^{(i)} = 0.2 \log n$
- $c_n^{(ii)} = 0.2n/\log n$
- $c_n^{(iii)} = 0.3\sqrt{n}$

Foi verificado que os resultados obtidos usando $c_n^{(i)}$ é inferior ao produzido usando $c_n^{(ii)}$ ou $c_n^{(iii)}$ conforme pode ser observado nas figuras 4.4, 4.5 e 4.6. O resultido obtido para o EDC usando $c_n^{(i)} = 0.2 \log n$ é inclusive inferior ao AIC (ver Figura 4.4). Verificou-se também que para as amostras analisadas neste trabalho, adotando $c_n = 0.2\sqrt{n}$ em (3.12) isso faz com que o EDC produza desempenho sempre semelhante ao apresentado pelo BIC para estes diferentes tamanhos de amostra (ver Figura 4.3). Na Figura 4.7 mostramos também o comportamento do EDC de acordo com os três $c_n^{'}s$ assumidos.

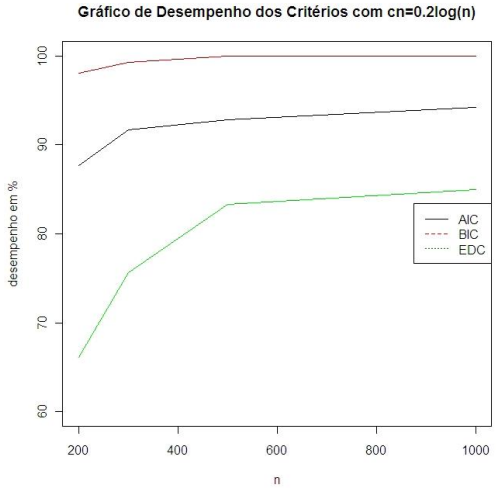

Figura 4.4: Desempenho do AIC,BIC e EDC com $c_n = 0.2\log(n)$.

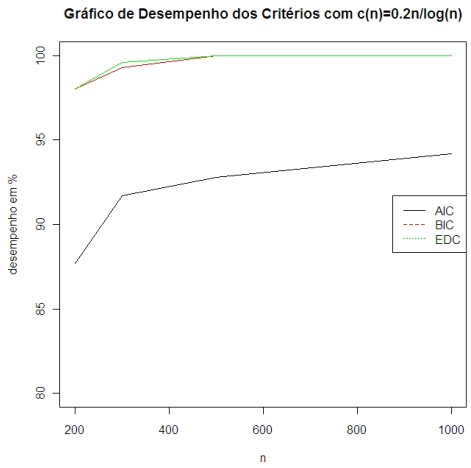

Figura 4.5: Desempenho do AIC,BIC e EDC com $c_n = 0.2n/logn$.

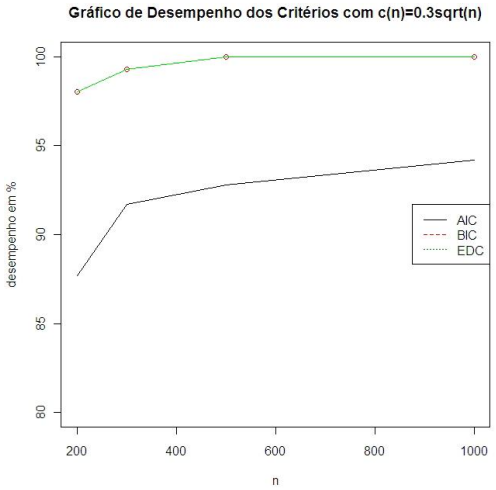

Figura 4.6: Desempenho do AIC,BIC e EDC com $c_n = 0.3\sqrt{n}$.

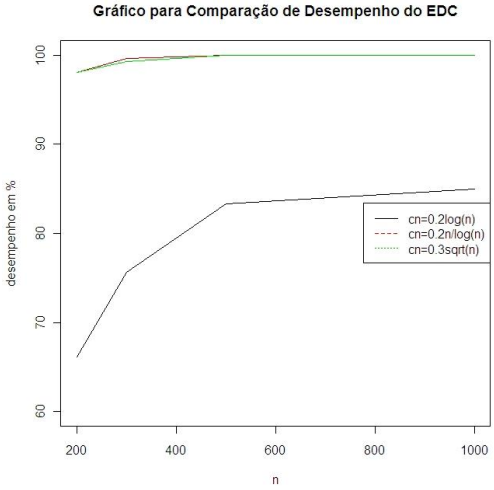

Figura 4.7: Desempenho do EDC para 3 diferentes $c_n^{'}s$.

4.3 Aplicação com Dados Reais

Nesta seção apresentamos a modelagem de dois conjuntos de dados amplamente conhecidos e discutidos na literatura estatística. O objetivo aqui é verificar se de acordo com a análise do histograma dos dados, os critérios nos dão bons indícios de quantas componentes podemos utilizar para realizarmos uma modelagem desses dados usando misturas finitas de densidades normais assimétricas.

O número de componentes indicados pelos critérios, bem como a modelagem para os referidos dados encontram-se nas subseções seguintes.

4.3.1 Os dados do PIB Per Capta

Nesta subseção utilizaremos o conjunto de dados já estudado anteriormente em Dias & Wedel (2004) referente ao PIB per capta em 1998 de 174 países. Para fazermos uma modelagem com o número de componentes sinalizado pelos critérios de informação empregados neste trabalho o procedimento adotado foi análogo ao empregado no estudo de simulação, tanto para inicialização do EM quanto para o emprego dos critérios. Vale lembrar que a função c_n usada aqui para o EDC foi $c_n = 0.2\sqrt{n}$.

O histograma desses dados sugere uma distribuição bimodal ou trimodal (ver Figura 4.11), o que em nossa abordagem pode ser entendido como o número de componentes para o modelo de mistura finita a ser adotado. Assim, os dados foram divididos em 2 e 3 grupos, respectivamente, e aplicamos o algoritmo EM nos modelos com os referidos números de componentes.

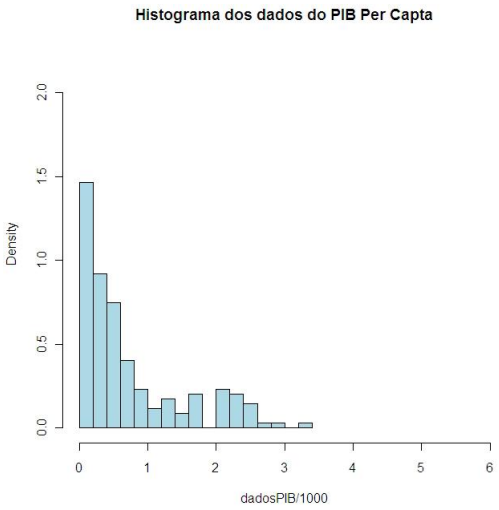

Figura 4.8: Histograma dos dados do PIB Per Capta.

Na Tabela 4.2 são apresentadas as log-verossimilhanças, os resultados dos critérios com o EDC assumindo $c_n = 0.2\sqrt{n}$ e o número de iterações para uma mistura com duas e três componentes, respectivamente. Através desta tabela, vemos que os critérios sinalizam que esta modelagem pode ser feita com 2 componentes. As estimativas dos parâmetros geradas pelo algoritmo EM e a densidade estimada podem ser vistas na Tabela 4.3 e na Figura 4.9, respectivamente. Pode ser visto que a mistura de densidades de duas componentes normais assimétricas para o modelo ajustado parece descrever de forma bem razoável os dados reais citados.

g	log-verossimilhança	AIC	BIC	EDC	Iterações
2	−94.95	205.24	227.35	214.32	7.960
3	−92.08	206.15	240.90	220.43	7.164

Tabela 4.2: Os critérios AIC,BIC e EDC nos dados do PIB, a log-verossimilhança estimada e o número de iterações.

Parâmetros	Estimativas	Parâmetros	Estimativas
ω_1	0.76	ω_2	0.24
ξ_1	0.05	ξ_2	1.92
σ_1^2	0.17	σ_2^2	0.32
λ_1	2432.89	λ_2	0.02

Tabela 4.3: Estimativas de máxima verossimilhança para os dados do PIB com 2 componentes SN.

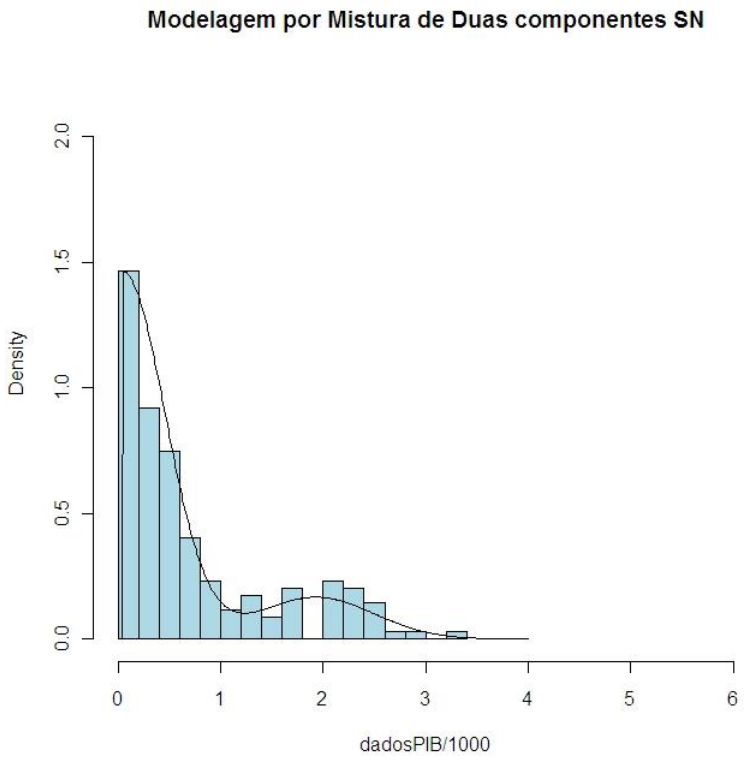

Figura 4.9: Densidade estimada para os dados do PIB com 2 componentes SN.

Para fins de comparação fizemos também a modelagem dos dados através de uma mistura de três componentes normais assimétricas, cujas estimativas dos parâmetros retornadas pelo algoritmo EM, estão na Tabela 4.4. A densidade estimada é apresentada na Figura 4.10 e nos dá indícios de que a modelagem realizada através da mistura de duas componentes descreve melhor os dados em detrimento da mistura com três componentes.

Parâmetros	Estimativas	Parâmetros	Estimativas	Parâmetros	Estimativas
ω_1	0.54	ω_2	0.2	ω_3	0.25
ξ_1	0.05	ξ_2	1.8	ξ_3	0.4
σ_1^2	0.05	σ_2^2	0.31	σ_3^2	0.19
λ_1	2136.63	λ_2	0.75	λ_3	7.94

Tabela 4.4: Estimativas dos parâmetros gerados pelo Algoritmo EM para 3 componentes.

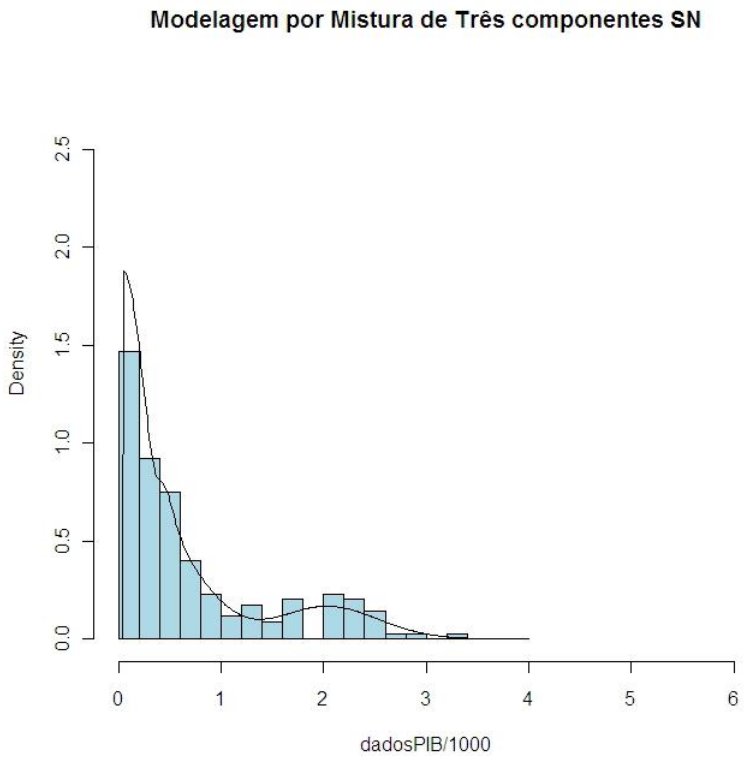

Figura 4.10: Densidade estimada para os dados do PIB com 3 componentes SN.

4.3.2 Os dados faithful

Consideremos nesta segunda aplicação, o conjunto de dados "Old Faithful Geyser"extraído de Silverman (1986). Estes dados se referem a 272 medições de tempo, em minutos, de erupção do géiser conhecido por *the old faithful* localizado no Parque Nacional de Yellowstone, Wyoming, EUA. Estes dados, como pode ser observado na Figura 4.11, aparentam ter comportamento assimétrico e bimodal. Por conta disso, procederemos de forma análoga à primeira aplicação para verificarmos se os critérios confirmam esta suposição.

Na Tabela 4.5 apresentamos as log-verossimilhanças, os resultados dos critérios com o EDC assumindo $c_n = 0.2\sqrt{n}$ e o número de iterações para uma mistura com duas e três componentes, respectivamente. Além disso, esta Tabela também mostra que os critérios novamente sinalizam que estes dados podem ser ajustados por uma mistura com duas componentes. As estimativas geradas pelo algoritmo

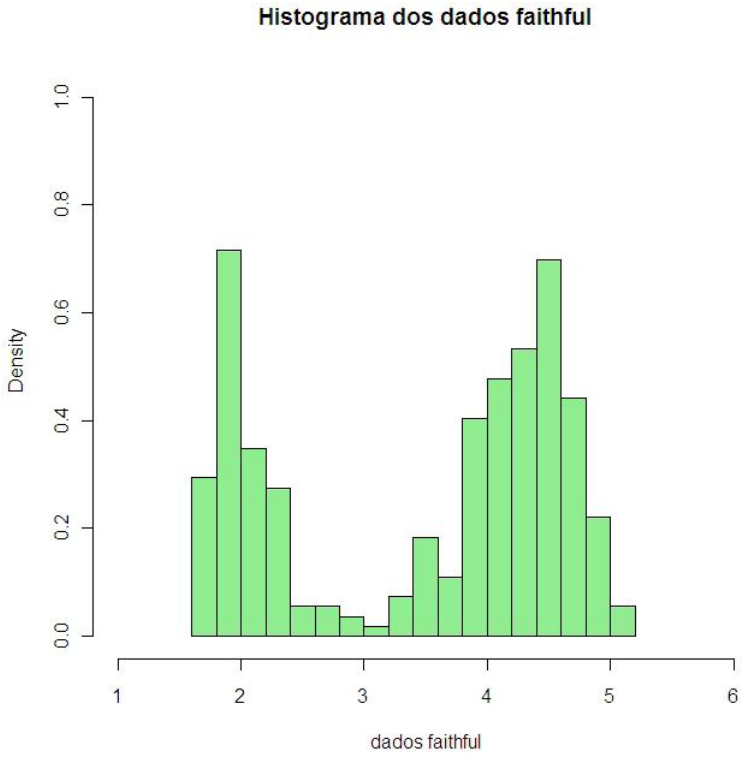

Figura 4.11: Histograma dos dados faithful geyser

EM estão na Tabela 4.6 e a densidade estimada pode ser vista na Figura 4.12. Também podemos observar, que o modelo ajustado por uma mistura de duas componentes normais assimétricas, visualmente indica que este modelo descreve razoavelmente estes dados.

g	log-verossimilhança	AIC	BIC	EDC	Iterações
2	−257.57	529.13	554.37	538.22	805
3	−257.26	536.52	576.19	550.81	1641

Tabela 4.5: Os critérios AIC,BIC e EDC nos dados Old Faithful, a log-verossimilhança estimada e o número de iterações.

Parâmetros	Estimativas	Parâmetros	Estimativas
ω_1	0.35	ω_2	0.65
ξ_1	1.73	ξ_2	0.47
σ_1^2	0.15	σ_2^2	4.80
λ_1	5.83	λ_2	-3.48

Tabela 4.6: Estimativas dos parâmetros para o modelo com 2 componentes SN.

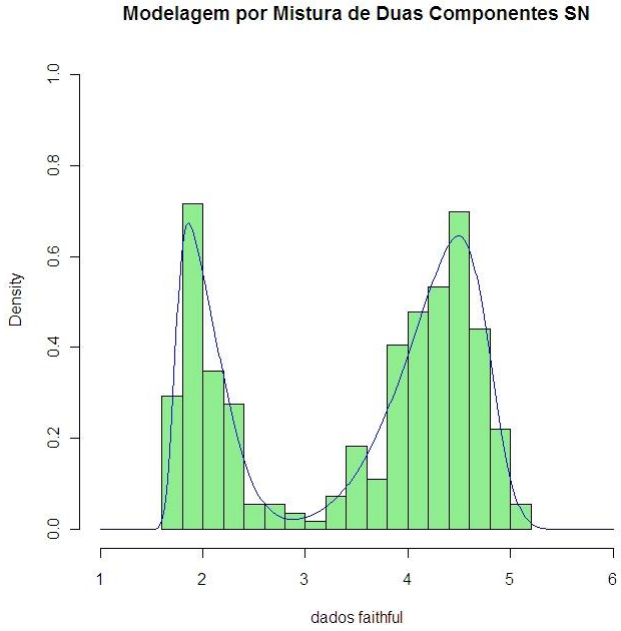

Figura 4.12: Densidade estimada para os dados faithful geyser com 2 componentes SN

Faremos também para este conjunto de dados, uma modelagem usando mistura de três componentes normais assimétricas conforme pode-se observar na Tabela 4.7 e na Figura 4.13

Parâmetros	Estimativas	Parâmetros	Estimativas	Parâmetros	Estimativas
ω_1	0.35	ω_2	0.5	ω_3	0.15
ξ_1	1.73	ξ_2	4.89	ξ_3	4.23
σ_1^2	0.14	σ_2^2	0.62	σ_3^2	0.08
λ_1	5.65	λ_2	-5.25	λ_3	-0.16

Tabela 4.7: Estimativas dos parâmetros gerados pelo Algoritmo EM para 3 componentes.

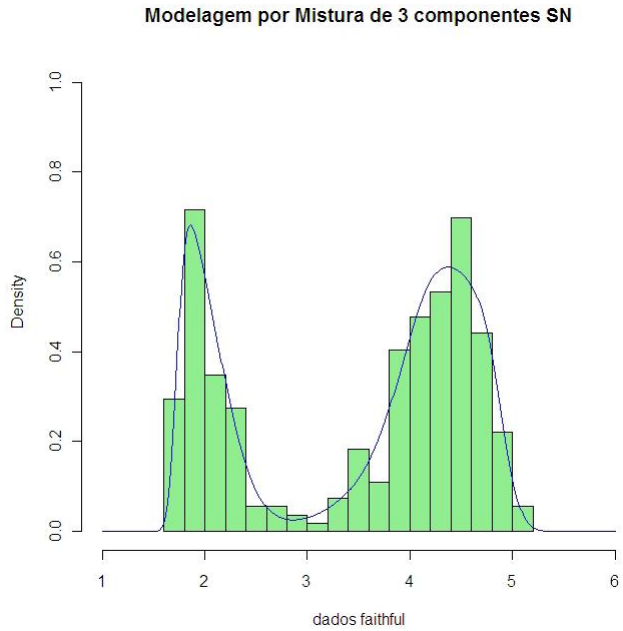

Figura 4.13: Densidade estimada para os dados faithful geyser com 3 componentes SN.

Capítulo 5

Conclusões

Apresentamos neste trabalho um estudo acerca do comportamento dos critérios AIC, BIC e EDC, com a finalidade de verificar qual deles se mostra mais apropriado para selecionar corretamente o número de componentes necessárias para fazermos uma modelagem de dados heterogêneos, dotados de assimetria, usando uma mistura finita de densidades normais assimétricas.

Com o emprego do algoritmo EM obtivemos as estimativas de máxima verossimilhança para os parâmetros do modelo o que tornou possível o emprego dos critérios de informação em estudo. Foi possível verificar que, embora o AIC ainda seja bastante utilizado na literatura em seleção de modelos, o seu desempenho foi inferior ao apresentado pelos demais. O fato de não levar em conta no seu termo de "penalização"o tamanho da amostra e sim apenas o número de parâmetros a ser estimado no modelo, torna o seu desempenho para os propósitos desse trabalho, pouco atraente.

O BIC como penaliza a log-verossimilhança mais fortemente que o AIC, por considerar o tamanho da amostra, apresentou desempenho superior. De acordo com este estudo, o BIC apresenta comportamento superior a noventa e cinco por cento de acerto a partir de amostras de tamanho 200. Para os propósitos desse trabalho esses resultados dão uma indicação de que podemos usá-lo para seleção de componentes em modelos de mistura finita de densidades normais assimétricas.

Quanto ao EDC, foi verificado que seu desempenho está diretamente relacionado à escolha do c_n. Isto é, pode ser tão bom quanto o BIC ou apresentar desempenho inclusive inferior ao AIC dependendo da escolha da função c_n. Um dado importante foi verificado ao longo desse trabalho: para os tamanhos de amostras analisados, o EDC empregado com $c_n = 0.2\sqrt{n}$ ou $c_n = 0.2n/log(n)$ não apresenta neste contexto de seleção de modelos, desempenho inferior ao AIC e BIC. A modelagem feita com dados reais vieram ratificar que a sinalização dos critérios para a escolha de quantas componentes usar numa modelagem por mistura finita de densidades, é bastante confiável.

Algumas dificuldades computacionais foram encontradas quando tentamos trabalhar com amostras de tamanho menor ou igual a 100 haja vista que o método utilizado para dividir esta amostra conforme descrito no Capítulo 4, o k-means, às vezes deixava grupos com muito poucas observações o que impediu a implementação computacional. Isto justifica o fato de termos usado amostras de tamanho maior ou igual a 200.

Devido a classe de funções que satisfazem as exigências da função c_n ser muito ampla, se torna inviável realizar uma investigação exaustiva nessa classe. Mas, podemos conjecturar de acordo com esse estudo, ser possível obter nesta classe uma nova função c_n, diferente das utilizadas neste trabalho e que apresente para amostras menores que 200 resultados tão bons quanto os já apresentados. Assim, um possível estudo seria avaliar outras funções, de forma a ampliar as funções c_n assumidas aqui e que contemple tamanhos de amostras diferentes das que foram objeto desse trabalho. Um outro estudo a ser feito seria estender este experimento para o caso multivariado.

Apêndice A

A Distribuição Normal Assimétrica

Apresentaremos aqui as demostrações de algumas das Propridades, Proposições e Lemas da distribuição normal assimétrica apresentadas no capítulo 1.

Propriedade 1.1.1 *A função densidade de uma variável aleatória* $Y \sim SN(0)$ *é idêntica à de uma variável aleatória* $X \sim N(0,1)$

Prova: Dado a densidade (1.1), com $\lambda = 0$, seque-se que

$$f(x;\lambda) \;=\; 2\phi(x)\,\Phi(0) \;=\; 2\phi(x)\,\tfrac{1}{2} \;=\; \phi(x)$$

onde $\phi(x) = N(0,1)$

Propriedade 1.1.2 *Quando* $\lambda \to \infty$ *a densidade (1.1) tende para* $2\phi(x)I_{x>o}$*, a qual corresponde à fdp de uma seminormal.*

Prova: Se $X \sim SN(\lambda)$ com fdp como em (1.1),
então

$$\Phi(\lambda\, x) = P(X \leq \lambda\, x)$$

logo, quando $\lambda \to \infty$

$$\Phi(\lambda\, x) = 1$$

ou seja, $lim_{\lambda\to\infty}\Phi(\lambda\, x) = 1$

Nessas condições, a densidade (1.1) resulta em, $f(x;\lambda) = 2\phi(x)$.

□

Proposição 1.1.1 *Se* $Y \sim N(0,1)$ *e* $Z \sim SN(\lambda)$ *então* $\mid Z \mid$ *e* $\mid Y \mid$ *tem a mesma função densidade de probabilidade.*

Prova:

Sejam $y > 0$ e $z > 0$,

$$\begin{aligned}
P(\mid Y \mid \leq y) &= P(-y \leq Y \leq y) \\
&= \Phi(y) - \Phi(-y) \\
&= 2\Phi(y) - 1. \\
P(\mid Z \mid \leq z) &= P(-z \leq Z \leq y) = F_Z(z;\lambda) - F_Z(-z;\lambda) \\
&= \Phi(z) - 2T(z;\lambda) - [\Phi(-z) - 2T(-z;\lambda)] \\
&= \Phi(z) - 2T(z;\lambda) - 1 + \Phi(z) + 2T(-z;\lambda) \\
&= 2\Phi(z) - 1.
\end{aligned}$$

□

Propriedade 1.0.6 $1 - F_Z(-z;\lambda) = F_Z(z;-\lambda)$

Prova:

$$\begin{aligned}
1 - F_Z(-z;\lambda) &= 1 - [\Phi(-z) - 2T(-z;\lambda)] \\
&= 1 - \Phi(-z) + 2T(-z;\lambda) \\
&= \Phi(z) + 2T(z;\lambda) \\
&= \Phi(z) - 2T(z;-\lambda) \\
&= F_Z(z;-\lambda)
\end{aligned}$$

Propriedade 1.0.7 Se $Z \sim SN(\lambda)$, então $-Z \sim SN(-\lambda)$

Prova:

Seja $W = -Z$; então,

$$\begin{aligned} P(W \leq w) &= P(-Z \leq -w) = P(Y > -w) = 1 - P(Z \leq -w) \\ &= 1 - F_Z(-w; \lambda) \\ &= F_Z(w; -\lambda) \end{aligned}$$

Assim,

$$W = -Z \sim SN(-\lambda).$$

Propriedade 1.0.8 $F_Z(z; 1) = [\Phi(z)]^2$.

Prova:

$$\begin{aligned} F_Z(z; 1) &= \Phi(z) - 2T(z; 1) = \Phi(z) - \Phi(z)\Phi(-z) \\ &= \Phi(z)[1 - \Phi(-z)] \\ &= [\Phi(z)]^2 \end{aligned}$$

Apêndice B

A Distribuição Normal Truncada

Neste apêndice definiremos o modelo normal truncada, sua função geradora de momentos e demonstração do Lema 1.1.1 extremamente útil no cálculo do passo E do algoritmo EM.

B.1 O Modelo Normal Truncada

Seja $X \sim N(\mu, \sigma^2)$. Se truncarmos os valores de X para valores em $A = (a_1, a_2)$,então uma variável aleatória $Y \sim X|X \in A$ tem distribuição normal truncada em A, denotada por $Y \sim NT(\mu, \sigma^2; (a_1, a_2))$ cuja função densidade de probabilidade é

$$f(y \mid \mu, \sigma^2) = \left\{ \Phi(\frac{a_2 - \mu}{\sigma}) - \Phi(\frac{a_1 - \mu}{\sigma}) \right\}^{-1} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}, a_1 < y < a_2 \tag{B.1}$$

A função Geradora de Momentos da Normal Truncada é dado como segue:

$$\begin{aligned} M(t) = E\left[e^{tY} \mid Y \in A\right] &= \frac{\int_{a_1}^{a_2} e^{ty} f(y) dy}{\Phi(\frac{a_2-\mu}{\sigma}) - \Phi(\frac{a_1-\mu}{\sigma})} \\ &= e^{\mu t + \sigma^2 t^2/2} \frac{\Phi(\frac{a_2-\mu}{\sigma} - \sigma t) - \Phi(\frac{a_1-\mu}{\sigma} - \sigma t)}{\Phi(\frac{a_2-\mu}{\sigma}) - \Phi(\frac{a_1-\mu}{\sigma})} \end{aligned} \tag{B.2}$$

A última igualdade segue de

$$\frac{1}{\sigma\sqrt{(2\pi)}} \int_{a_1}^{a_2} e^{ty} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy = \frac{1}{\sigma\sqrt{(2\pi)}} \int_{a_1}^{a_2} e^{-\frac{1}{2\sigma^2}[y-(\sigma^2 t+\mu)]^2 - (\sigma^2 t+\mu)^2 + \mu^2} dy$$

$$
\begin{aligned}
&= e^{-\frac{1}{2\sigma^2}[\mu^2-(\sigma^2t+\mu)^2]}\frac{1}{\sigma\sqrt{2\pi}}\int_{a_1}^{a_2} e^{-\frac{1}{2}\frac{(y-\gamma}{\sigma})^2}dy \\
&= e^{\mu t+\sigma^2t^2}\int_{a_1}^{a_2}\frac{1}{\sigma}\phi\big(\frac{y-\gamma}{\sigma}\big)dy \\
&= e^{\mu t+\sigma^2t^2/2}\Big[\Phi\big(\frac{a_2-\gamma}{\sigma}\big)-\Phi\big(\frac{a_1-\gamma}{\sigma}\big)\Big].
\end{aligned}
$$

onde $\gamma = \sigma^2 t + \mu$.

Lema 1.1.1 *Seja* $X \sim NT(\mu,\sigma^2)I\{a_1 < x < a_2\}$ *uma distribuição normal truncada com densidade dada por (B.1)*,Então

$$
\begin{aligned}
(i)E(X) &= \mu - \sigma\frac{\phi(\alpha_2)-\phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)} && \text{(B.3)}\\
(ii)E(X^2) &= \mu^2+\sigma^2-\sigma^2\frac{\alpha_2\phi(\alpha_2)-\phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)} - 2\mu\sigma\frac{\phi(\alpha_2)-\phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)} && \text{(B.4)}
\end{aligned}
$$

Prova:

Derivando (B.2) com respeito a t, temos que

$$
\begin{aligned}
M'(t) &= \mu+\sigma^2te^{\mu t+\sigma^2t^2/2}\left[\frac{\Phi\big(\frac{a_2-\mu}{\sigma}-\sigma t\big)-\Phi\big(\frac{a_1-\mu}{\sigma}-\sigma t\big)}{\Phi\big(\frac{a_2-\mu}{\sigma}\big)-\Phi\big(\frac{a_1-\mu}{\sigma}\big)}\right]+ \\
&\quad +e^{\mu t+\sigma^2t^2/2}\left[\sigma\frac{\phi\big(\frac{a_2-\mu}{\sigma}-\sigma t\big)-\phi\big(\frac{a_1-\mu}{\sigma}-\sigma t\big)}{\Phi\big(\frac{a_2-\mu}{\sigma}\big)-\Phi\big(\frac{a_1-\mu}{\sigma}\big)}\right] \qquad \text{(B.5)}
\end{aligned}
$$

logo,

$(i)E[X \mid X \in A] = M'(t)\mid_{t=0} = \mu - \sigma\frac{\phi(\alpha_2)-\phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)}$.

□

De forma análoga,

$$
\begin{aligned}
M''(t) &= \Big(\mu e^{\mu t+\frac{\sigma^2t^2}{2}}+\sigma^2te^{\mu t+\frac{\sigma^2t^2}{2}}\Big)\left[\frac{\Phi(\alpha_2-\sigma t)-\Phi(\alpha_1-\sigma t)}{\Phi(\alpha_2)-\Phi(\alpha_1)}\right]- \\
&\quad -e^{\mu t+\frac{\sigma^2t^2}{2}}\left[\sigma\frac{\phi(\alpha_2-\sigma t)-\phi(\alpha_1-\sigma t)}{\Phi(\alpha_2)-\Phi(\alpha_1)}\right] \\
&= \mu(\mu+\sigma^2t)e^{\mu t+\frac{\sigma^2t^2}{2}}+\sigma^2\Big[e^{\mu t+\frac{\sigma^2t^2}{2}}+ \\
&\quad +t(\mu+\sigma^2t)e^{\mu t+\frac{\sigma^2t^2}{2}}\Big]\left[\frac{\Phi(\alpha_2-\sigma t)-\Phi(\alpha_1-\sigma t)}{\Phi(\alpha_2)-\Phi(\alpha_1)}\right]+ \\
&\quad +\Big(\mu e^{\mu t+\frac{\sigma^2t^2}{2}}+\sigma^2te^{\mu t+\frac{\sigma^2t^2}{2}}\Big)(-\sigma)\left[\frac{\phi(\alpha_2-\sigma t)-\phi(\alpha_1-\sigma t)}{\Phi(\alpha_2)-\Phi(\alpha_1)}\right]-
\end{aligned}
$$

$$-(\mu+\sigma^2 t)e^{\mu t+\frac{\sigma^2 t^2}{2}}\left[\sigma\frac{\phi(\alpha_2-\sigma t)-\phi(\alpha_1-\sigma t)}{\Phi(\alpha_2)-\Phi(\alpha_1)}\right]+$$
$$+\sigma^2\frac{\phi'(\alpha_2-\sigma t)-\phi'(\alpha_1-\sigma t)}{\Phi(\alpha_2)-\Phi(\alpha_1)}$$

logo,

$$\begin{aligned}
(ii)E[X^2 \mid X\in A] &= \\
M''(t)\mid_{t=0} &= \mu^2+\sigma^2\frac{\Phi(\alpha_2)-\Phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)}-\mu\sigma\frac{\phi(\alpha_2)-\phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)}- \\
&\quad -\mu\sigma\frac{\phi(\alpha_2)-\phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)} \\
&= \sigma^2+\mu^2+\sigma^2\frac{\phi'(\alpha_2)-\phi'(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)}-2\mu\sigma\frac{\phi(\alpha_2)-\phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)} \\
&= \mu^2+\sigma^2-\sigma^2\frac{\alpha_2\phi(\alpha_2)-\alpha_1\phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)}-2\mu\sigma\frac{\phi(\alpha_2)-\phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)}
\end{aligned}$$

□

onde,

$$\begin{aligned}
\phi'(\alpha_i) &= -\Big(\frac{a_i-\mu}{\sigma}\Big)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{a_i-\mu}{\sigma})^2} \\
&= -\alpha_i\phi(\alpha_i)
\end{aligned}$$

Apêndice C

O algoritmo EM

Como um algoritmo tipo EM foi a ferramenta utilizada para encontramos as estimativas de máxima verossimilhança no decorrer deste trabalho, apresentamos aqui um breve resumo dessa teoria que pode ser ampliada em McLachlan e Pell (2000).

C.1 Teoria

O algoritmo EM desenvolvido por Dempester et al. (1977) é um método iterativo que no contexto de modelo de mistura finita é extremamente eficaz para dar solução a equação de máxima verossimilhança, gerando portanto as estimativas de máxima verossimilhança.

Esta teoria de estimação por máxima verossimilhança acontece em dois passos: o passo E e o passo M. No passo E são calculadas as esperanças condicionais da função de log-verossimilhança dos completos dadas as observações, conhecida como função Q. Maximizando esta função obtemos as estimativas de máxima verossimilhança atualizadas, o que denominamos de passo M. Este processo é então repetido até que se atinja uma condição de parada pré-estabelecida. Para a inicialização do algoritmo é necessário que seja fornecido um valor inicial, o qual chamaremos de $\theta^{(0)}$.

Em geral o algoritmo EM é definido de forma que se encontre uma sequência $\{\Theta^{(k)}\}$ que converge para um ponto estacionário de $l_c(\theta)$. A partir de um valor $\theta^{(k)}$ gerado pelo algoritmo sua atualização $\theta^{(k+1)}$ acontece da seguinte forma:
i) Passo E: Calcula-se as esperanças condicionais;
ii)Passo M: escolhe-se $\theta^{(k+1)}$ a partir de um conjunto de valores de $\theta \in \Omega$ que maximizam $Q(\theta|\theta^{(k)})$ sobre o espaço paramétrico Ω. Ou seja, $\theta^{(k+1)} \in \arg max_{\theta\in\Omega} Q(\theta|\theta^{(k)})$.

iii) Condição de Parada: as iterações do algoritmo são feitas através da repetição alternada dos passo E e M até que um critério de convergência seja atingido. É comum utilizar a diferença $(l_c(\Theta^{(k+1)}) - l_c(\Theta^{(k)})$ o que significa que a convergência será atingida quando esta diferença for menor que uma constante c especificada, ou seja

$$(l_c(\Theta^{(k+1)}) - l_c(\Theta^{(k)}) < c \tag{C.1}$$

Como as estimativas geradas pelo EM é uma sequência monótona não-decrescente segue-se que

$$l_c(\Theta^{(k+1)}) > l_c(\Theta^{(k)}) \tag{C.2}$$

Apêndice D

Distribuição Normal Assimétrica: Estimação dos Parâmetros via Algoritmo tipo EM

Neste apêndice mostramos os cálculos necessários para realizarmos os passos E e CM do algoritmo ECM. Isto inclui a determinação da função Q e as derivadas de Q com relação aos parâmetros do modelo a serem estimados.

D.1 Função Q

$$
\begin{aligned}
Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(k)}) &= E_{\boldsymbol{\Theta}}\Big[l_c(\boldsymbol{\Theta}) \mid y_i\Big] \\
&= -\frac{n}{2}E(\log(\Gamma)|y_i) - \frac{1}{2\Gamma}E\left(\sum_{i=1}^{n}(y_i^2+\xi^2+(\Delta t_i)^2-2\Delta t_i y_i - 2\xi y_i + 2\Delta t_i \xi)|y_i)\right) \\
&= -\frac{n}{2}\log(\Gamma) - \frac{1}{2\Gamma}\sum_{i=1}^{n}E(y_i^2|y_i) - \frac{1}{2\Gamma}\sum_{i=1}^{n}E(\xi^2|y_i) - \frac{1}{2\Gamma}\sum_{i=1}^{n}E[(\Delta t_i)^2|y_i] + \\
&\quad + \frac{1}{\Gamma}\sum_{i=1}^{n}E(\Delta t_i y_i|y_i) + \frac{1}{\Gamma}\sum_{i=1}^{n}E(\xi y_i|y_i) - \frac{1}{\Gamma}\sum_{i=1}^{n}E(\xi \Delta t_i|y_i) \\
&= -\frac{n}{2}\log(\Gamma) - \frac{1}{2\Gamma}\sum_{i=1}^{n}y_i^2 - \frac{1}{2\Gamma}\sum_{i=1}^{n}\xi^2 - \frac{1}{2\Gamma}\sum_{i=1}^{n}\Delta^2\overbrace{E(t_i^2|y_i)}^{s_{2i}} + \frac{1}{\Gamma}\sum_{i=1}^{n}\Delta y_i\overbrace{E(t_i|y_i)}^{s_{1i}} + \\
&\quad + \frac{1}{\Gamma}\sum_{i=1}^{n}\xi y_i - \frac{1}{\Gamma}\sum_{i=1}^{n}\xi\Delta\overbrace{E(t_i|y_i)}^{s_{1_i}} \\
&= -\frac{n}{2}\log(\Gamma) - \frac{1}{2\Gamma}\sum_{i=1}^{n}y_i^2 - \frac{1}{2\Gamma}(n\Delta^2\sum_{i=1}^{n}s_{2i}) + \frac{2}{2\Gamma}(n\Delta\sum_{i=1}^{n}(y_i-\xi)s_{1i}) +
\end{aligned}
$$

$$+\frac{2}{2\Gamma}(n\xi\sum_{i=1}^{n}y_i)-\frac{n\xi^2}{2\Gamma} \tag{D.1}$$

Calculando as derivadas parciais de (D.1) com respeito a $\hat{\xi}^{(k)}, \hat{\Delta}^{(k)}$, $\hat{\Gamma}^{(k)}$ e igualando a zero, temos:

$$\begin{aligned}\left.\frac{\partial Q(\boldsymbol{\Theta}\mid\boldsymbol{\Theta}^{(k)})}{\partial\hat{\xi}^{(k)}}\right|_{\hat{\xi}^{(k)}=\hat{\xi}^{(k+1)}}=0 \quad &\Rightarrow \quad -\frac{1}{\Gamma}n\Delta\sum_{i=1}^{n}\hat{s}_{1i}+\frac{1}{\Gamma}n\sum_{i=1}^{n}y_i-\frac{1}{\Gamma}n\hat{\xi}^{(k+1)}=0\\ &\Rightarrow \quad \hat{\xi}^{(k+1)}=\sum_{i=1}^{n}(y_i-\Delta\hat{s}_{1i})\end{aligned} \tag{D.2}$$

$$\begin{aligned}\left.\frac{\partial Q(\boldsymbol{\Theta}\mid\boldsymbol{\Theta}^{(k)})}{\partial\hat{\Delta}^{(k)}}\right|_{\hat{\Delta}^{(k)}=\hat{\Delta}^{(k+1)}}=0 \quad &\Rightarrow \quad -\frac{1}{\Gamma}n\Delta\sum_{i=1}^{n}\hat{s}_{2i}-\frac{1}{\Gamma}n\sum_{i=1}^{n}(y_i-\hat{\xi}^{(k+1)})\hat{s}_{1i}=0\\ &\Rightarrow \quad \hat{\Delta}^{k+1}=\frac{\sum_{i=1}^{n}(y_i-\hat{\xi}^{(k+1)})\hat{s}_{1i}}{\sum_{i=1}^{n}\hat{s}_{2i}}\end{aligned} \tag{D.3}$$

$$\begin{aligned}\left.\frac{\partial Q(\boldsymbol{\Theta}\mid\boldsymbol{\Theta}^{(k)})}{\partial\hat{\Gamma}^{(k)}}\right|_{\hat{\Gamma}^{(k)}=\hat{\Gamma}^{(k+1)}}=0 \Rightarrow -\frac{n}{2\hat{\Gamma}^{(k+1)}}+\frac{2\sum_{i=1}^{n}y_i^2}{2\hat{\Gamma}^{2^{(k+1)}}}+\frac{n\hat{\Delta}^{2^{(k+1)}}\sum_{i=1}^{n}\hat{s}_{2i}}{2\hat{\Gamma}^{2^{(k+1)}}}-\\ -\frac{2n\hat{\Delta}^{(k+1)}\sum_{i=1}^{n}(y_i-\hat{\xi}^{(k+1)})\hat{s}_{1i}}{2\hat{\Gamma}^{2^{(k+1)}}}-\frac{2n\hat{\xi}^{(k+1)}\sum_{i=1}^{n}y_i}{2\hat{\Gamma}^{2^{(k+1)}}}+\frac{n\hat{\xi}^{2^{(k+1)}}}{2\hat{\Gamma}^{2^{(k+1)}}}=0\\ \Rightarrow n\hat{\Gamma}^{(k+1)}=\sum_{i=1}^{n}y_i^2+n\hat{\Delta}^{2^{(k+1)}}\sum_{i=1}^{n}\hat{s}_{2i}-2n\hat{\Delta}^{(k+1)}\sum_{i=1}^{n}(y_i-\hat{\xi}^{(k+1)})\hat{s}_{1i}-\\ -2n\hat{\xi}^{(k+1)}\sum_{i=1}^{n}y_i+n\hat{\xi}^{2^{(k+1)}}\\ \Rightarrow n\hat{\Gamma}^{(k+1)}=\sum_{i=1}^{n}y_i^2+\sum_{i=1}^{n}\hat{\Delta}^{2^{(k+1)}}\hat{s}_{2i}-2\sum_{i=1}^{n}\hat{\Delta}^{(k+1)}(y_i-\hat{\xi}^{(k+1)})\hat{s}_{1i}-\\ -2\sum_{i=1}^{n}\hat{\xi}^{(k+1)}y_i+\sum_{i=1}^{n}\hat{\xi}^{2^{(k+1)}}\\ \hat{\Gamma}^{(k+1)}=\frac{\sum_{i=1}^{n}(y_i^2+2\hat{\xi}^{(k+1)}y_i-\hat{\xi}^{2(k+1)})-2\sum_{i=1}^{n}\hat{\Delta}^{(k+1)}(y_i-\hat{\xi}^{(k+1)})\hat{s}_{1i}+\sum_{i=1}^{n}\hat{\Delta}^{2(k+1)}}{n}\\ \hat{\Gamma}^{(k+1)}=\frac{\sum_{i=1}^{n}\left[(y_i-\hat{\xi}^{(k+1)})^2-2(y_i-\hat{\xi}^{(k+1)})\hat{\Delta}^{(k+1)}\hat{s}_{1i}-\hat{\Delta}^{2^{(k+1)}}\hat{s}_{2i}\right]}{n}\end{aligned} \tag{D.4}$$

D.2 Algoritmo ECM para Misturas Finitas de Normais Assimétricas

D.2.1 A função Q e as atualizações para o algoritmo ECM

$$
\begin{aligned}
Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(k)}) &= E_{\boldsymbol{\Theta}}\Big[l_c(\boldsymbol{\Theta}) \mid y_i\Big] \\
&= E_{\boldsymbol{\Theta}}\Big[\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}(\log\omega_j \mid y_i) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}(\log\Gamma \mid y_i) - \\
&\quad -\frac{1}{2\Gamma}\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}((y_i - \xi_j - \Delta_j t_i^2) \mid y_i)\Big] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\log\omega_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\log\Gamma_j - \\
&\quad -\frac{1}{2\Gamma_j}\sum_{i=1}^{n}\sum_{j=1}^{k} E\Big[Z_{ij}(y_i^2 \mid y_i + \xi_j^2 \mid y_i + \Delta_j^2 t_i^2 \mid y_i - 2y_i\xi_j \mid y + \\
&\quad +2\xi_j\Delta_j t_i \mid y_i - 2y_i\Delta_j t - i \mid y_i)\Big] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\log\omega_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\log\Gamma_j - \frac{1}{2\Gamma_j}\sum_{i=1}^{n}\sum_{j=1}^{k} E\Big[Z_{ij}y_i^2\Big] - \\
&\quad -\frac{1}{2\Gamma_j}\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\xi_j^2 - \frac{1}{2\Gamma_j}\sum_{i=1}^{n}\sum_{j=1}^{k} E(Z_{ij}t_i^2\Delta_j^2 \mid y_i) + \\
&\quad +\frac{\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}y_i\xi_i}{\Gamma_j} - \frac{1}{\Gamma}\sum_{i=1}^{n}\sum_{j=1}^{k} E(Z_{ij}t_i\xi_j\Delta_j \mid y_i) + \\
&\quad +\frac{1}{\Gamma_j}\sum_{i=1}^{n}\sum_{j=1}^{k} E(z_{ij}t_i\Delta_j y_i \mid y_i) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\log\omega_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\log\Gamma_j - \frac{1}{2\Gamma_j}\sum_{i=1}^{n}\sum_{j=1}^{k} EZ_{ij}y_i^2 - \\
&\quad -\frac{1}{2\Gamma_j}\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\xi_j^2 - \frac{1}{2\Gamma}\sum_{i=1}^{n}\sum_{j=1}^{k} \Delta_j^2\hat{s}_{3ij} + \frac{1}{\Gamma}\sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}y_i\xi_j - \\
&\quad -\frac{1}{\Gamma}\sum_{i=1}^{n}\sum_{j=1}^{k} \xi_j\Delta_j\hat{s}_{2ij} + \frac{1}{\Gamma}\sum_{i=1}^{n}\sum_{j=1}^{k} y_i\Delta_j\hat{s}_{2ij}
\end{aligned} \tag{D.5}
$$

Agora, calculando as derivadas parciais de (2.20) com respeito a $\hat{\xi}_j^{(k)}, \hat{\Delta}_j^{(k)}, \hat{\Gamma}_j^{(k)}$ e igualando a zero, temos:

$$
\begin{aligned}
\left.\frac{\partial Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(k)})}{\partial\hat{\xi}_j^{(k)}}\right|_{\hat{\xi}^{(k)}=\hat{\xi}^{(k+1)}} = 0 \quad &\Rightarrow \quad -\frac{\hat{\xi}_j^{(k+1)}\sum_{i=1}^{n}\hat{z}_{ij}}{\hat{\Gamma}_j} + \sum_{i=1}^{n}\hat{s}_{1ij}y_i - \sum_{i=1}^{n}\Delta_j^{(k)}\hat{s}_{2ij} = 0 \\
&\Rightarrow \quad \hat{\xi}_j^{(k+1)}\sum_{i=1}^{n}\hat{z}_{ij} = \sum_{i=1}^{n}\hat{s}_{1ij}y_i - \hat{\Delta}_j^{(k)}\hat{s}_{2ij}
\end{aligned}
$$

$$\Rightarrow \quad \hat{\xi}_j^{(k+1)} = \frac{\sum_{i=1}^n (\hat{s}_{1ij} y_i - \hat{\Delta}_j^{(k)} \hat{s}_{2ij})}{\sum_{i=1}^n \hat{s}_{1ij}} \tag{D.6}$$

$$\left.\frac{\partial Q(\boldsymbol{\Theta} \mid \hat{\boldsymbol{\Theta}}^{(k)})}{\partial \hat{\Delta}_j^{(k)}}\right|_{\hat{\Delta}^{(k)}=\hat{\Delta}^{(k+1)}} = 0 \quad \Rightarrow \quad -\hat{\Delta}_j^{(k+1)} \sum_{i=1}^n \hat{s}_{2ij} - \sum_{i=1}^n \hat{\xi}_j^{(k+1)} \hat{s}_{1ij} + \sum_{i=1}^n y_i \hat{s}_{1ij=0}$$

$$\Rightarrow \quad \hat{\Delta}_j^{(k+1)} = \frac{\sum_{i=1}^n (y_i - \xi_j^{(k+1)}) \hat{s}_{1ij}}{\sum_{i=1}^n \hat{s}_{2ij}} \tag{D.7}$$

$$\left.\frac{\partial Q(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)})}{\partial \hat{\Gamma}_j^{(k)}}\right|_{\hat{\Gamma}^{(k)}=\hat{\Gamma}^{(k+1)}} = 0 \Rightarrow$$

$$\begin{aligned}
\Rightarrow \quad & -\frac{1}{2} \sum_{i=1}^n \hat{z}_{ij} \frac{1}{\hat{\Gamma}_j} + \frac{1}{2\hat{\Gamma}^2} \sum_{i=1}^n \hat{z}_{ij} y_i^2 - \\
& - \frac{1}{2\hat{\Gamma}_j^{2^{(k+1)}}} \sum_{i=1}^n \hat{z}_{ij} \hat{\xi}_j^{2^{(k+1)}} - \frac{1}{2\hat{\Gamma}^2} \sum_{i=1}^n \hat{\Delta}_j^{2^{(k+1)}} \hat{s}_{2ij} + \\
& + \frac{1}{\hat{\Gamma}_j^{2^{(k+1)}}} \sum_{i=1}^n \hat{z}_{ij} y_i \hat{\xi}_j^{(k+1)} - \frac{1}{\hat{\Gamma}^2} \sum_{i=1}^n \hat{\xi}_j^{(k+1)} \hat{\Delta}_j^{(k+1)} \hat{s}_{2ij} + \\
& + \frac{1}{\hat{\Gamma}_j^{2^{(k+1)}}} \sum_{i=1}^n y_i \hat{\Delta}_j^{(k+1)} \hat{s}_{1ij} = 0 \\
\Rightarrow \quad & -\hat{\Gamma}_j^{(k+1)} \sum_{i=1}^n \hat{z}_{ij} + \sum_{i=1}^n \hat{z}_{ij} y_i^2 + \sum_{i=1}^n \hat{z}_{ij} \hat{\xi}_j^{2^{(k+1)}} + \\
& + \sum_{i=1}^n \hat{\Delta}_j^{2^{(k+1)}} \hat{s}_{2ij} - 2 \sum_{i=1}^n \hat{z}_{ij} y_i \hat{\xi}_j^{(k+1)} + \\
& + 2 \sum_{i=1}^n \hat{\xi}_j^{(k+1)} \hat{\Delta}_j^{(k+1)} \hat{s}_{1ij} - 2 \sum_{i=1}^n y_i \hat{\Delta}_j^{(k+1)} \hat{s}_{1ij} = 0 \\
\Rightarrow \quad & \hat{\Gamma}_j^{(k+1)} = \frac{\sum_{i=1}^n \hat{z}_{ij} (y_i - \hat{\xi}_j^{(k+1)})^2 - 2 \sum_{i=1}^n (y_i - \hat{\xi}_j^{(k+1)}) \hat{\Delta}_j^{(k+1)} \hat{s}_{1ij} + \sum_{i=1}^n \hat{\Delta}_j^{2^{(k+1)}} \hat{s}_{2ij}}{\sum_{i=1}^n \hat{z}_{ij}} \\
\Rightarrow \quad & \hat{\Gamma}_j^{(k+1)} = \frac{\sum_{i=1}^n (\hat{z}_{ij} (y_i - \hat{\xi}_j^{(k+1)})^2 - 2(y_i - \hat{\xi}_j^{(k)}) \hat{\Delta}_j^{(k+1)} \hat{s}_{1ij} + \hat{\Delta}_j^{2^{(k+1)}} \hat{s}_{2ij})}{\sum_{i=1}^n \hat{z}_{ij}}
\end{aligned} \tag{D.8}$$

Referências Bibliográficas

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**(6), 716–723.

Arellano-Valle, R., Bolfarine, H. & Lachos, V. (2005). Skew-normal linear mixed models. *Journal of Data Science*, **3**, 415–438.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171–178.

Azzalini, A. (1986). Further results on a class of distribution which includes the normal ones. *Statistica*, **46**, 199–208.

Azzalini, A. & Capitanio, A. (1999). Statistical applications of the multivariate skew-normal distribution . *J. Roy.Statist. Soc. Ser. B*, **61**, 579–602.

Azzalini, A. & Dalla Vale, A. (1996). The Multivariate Skew-Normal distribution . *Biometrika*, **83**, 715–726.

Bai, Z. D., Krishnaiah, P. R. & Zhao, L. C. (1989). On rates of convergence of efficient detection criteria in signalprocessing with white noise. *IEEE Transactions On Information Theory*, **35**(2), 380–388.

Basso, R. M. (2009). Misturas finitas de escalas skew normal . *Dissertação de Mestrado-UNICAMP*.

Bayes, C. L. & Branco, M. D. (2007). Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Braz. J. Probab. Statist*.

Biernacki, C., Celeux, G. & Govaert, G. (2000). Assessing a mixture model for clustering with the integratedcompleted likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 719–725.

Celeux, G. & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, **13**(2), 195–212.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Dias, J. & Wedel, M. (2004). An empirical comparison of EM, SEM and MCMC performance for problematic Gaussian mixture likelihoods. *Statistics and Computing*, **14**(4), 323–332.

Henze, N. (1986). A probabilistic representation of the skew-normal distribution. *Scandinavian Journal of Statistics*, **13**, 271–275.

Johnson, N. L., Kotz, S. & Balakrisnan, N. (1994). *Continuous Univariate Distributions*. Wiley, London.

Kass, R. & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430).

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86.

Lin, T. I., Lee, J. C. & Yen, S. Y. (2007). Finite Mixture Modelling Using the Skew Normal Distribution. *Statistica Sinica*, **17**, 909–927.

McLachlan, G. J. & Pell, D. (2000). *Finite Mixture Models*. Wiely, New York.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall/CRC.

Wichern, D. & Johnson, R. (2007). *Applied multivariate statistical analysis*. Pearson Prentice Hall.

Zhao, L., Dorea, C. C. Y. & Gonçalves, C. R. (2001). On Determination of the Order of a Markov Chain. *Statistical Inference for Stochastic Processes*, **4**, 273–282.